GigaScience

# KOREF_S1: the phased, parental Trio-binned Korean reference genome using long-reads and Hi-C sequencing methods

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | GIGA-D-21-00219R2 |
| **Full Title:** | KOREF_S1: the phased, parental Trio-binned Korean reference genome using long-reads and Hi-C sequencing methods |
| **Article Type:** | Data Note |
| **Funding Information:** | ministry of smes and startups (P0016193) — Dr. Jong Hwa Bhak |
| **Abstract:** | Background<br><br>KOREF is the Korean reference genome which was constructed with various sequencing technologies including long reads, short reads, and optical mapping methods. It is also the first East Asian multiomic reference genome accompanied by extensive clinical information, time series and multiomic data, and his parental sequencing data. However, it was still not a chromosome-scale reference. Here, we updated the previous KOREF assembly to a new chromosome-level haploid assembly of KOREF, KOREF_S1v2.1. ONT PromethION, PacBio Hifi-CCS, and Hi-C technology were used to build the most accurate East Asian reference assembled so far.<br><br>Results<br><br>We produced 705 Gb ONT reads and 114 Gb PacBio HiFi reads, and corrected ONT reads by PacBio reads. The corrected ultra-long reads reached higher accuracy of 1.4% base-errors than the previous KOREF_S1v1.0, which was mainly built with short reads. KOREF has parental genome information, and we successfully phased it using a trio-binning method acquiring a near-complete haploid-assembly. The final assembly resulted in total length of 2.9 Gb with an N50 of 150 Mb, and the longest scaffold covered 97.3% of GRCh38's chromosome 2. And the final assembly showed high base accuracy, less than 0.01% of base-errors.<br><br>Conclusions<br><br>KOREF_S1v2.1 is the first chromosome-scale haploid assembly of the Korean reference genome with high contiguity and accuracy. Our study provides useful resources of the Korean reference genome and demonstrates a new strategy of hybrid assembly which collaborates ONT's PromethION and PacBio's HiFi-CCS. |
| **Corresponding Author:** | Jong Hwa Bhak, Ph.D.<br>UNIST<br>Ulsan, Ulsan KOREA, REPUBLIC OF |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | UNIST |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Hui-su Kim, Ph.D. |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Hui-su Kim, Ph.D.<br>Sungwon Jeon, Ph.D.<br>Yeonkyung Kim<br>Changjae Kim, Ph.D. |

| | |
|---|---|
| | Jihun Bhak |
| | Jong Hwa Bhak, Ph.D. |
| **Order of Authors Secondary Information:** | |
| **Response to Reviewers:** | Revision for KOREF manuscript to the GigaScience<br><br>Reviewer #1:<br>The authors present an improved reference assembly for an extensively characterized Korean son in a trio. Specifically, they partition ONT and HiFi reads by haplotype, correct ONT reads with HiFi reads, and assemble the corrected reads followed by scaffolding with Hi-C. This is an assembly approach I haven't seen before, and it yields impressive chromosome-scale scaffolds. However, the completeness, contig N50's, and QV's are substantially worse than recent assemblies from HiFi data alone, particularly from trio-hifiasm,<br><br>We noted and emphasized the limitation of our assembly.<br><br>so I think the authors need to better emphasize the limitations of their assembly, as well as its strengths. If this is made clear, I expect this to be a useful manuscript.<br><br>1. The authors should clearly state in the results that their QV of ~44 is substantially lower than the QV of ~50 for recently published hifiasm assemblies that use HiFi data alone, albeit HiFi with longer read lengths (https://www.nature.com/articles/s41592-020-01056-5/tables/3)<br><br>Thank you. It is true that our QV was substantially lower than the QV of Hifiasm assemblies. We have now additionally compared contig assemblies of KOREF, HG00733, and HG002. All results are in Table 5. HG002 assembly showed highest QV of 51.6 and PromethION assembly of KOREF showed lowest QV of 33.8. HiFi-PromethION hybrid assembly of KOREF scored higher QV (42.2) against PromethION assembly. However, it was lower than the HiFi assembly of KOREF (QV 45.1). We noted this on Line 303.<br><br>2. The authors should clearly state in the main text that their assembly's completeness in Table S3 is only 90-92%, >10x more missing sequence than their hifiasm assembly (99.2-99.7%)<br><br>We agree. The hifiasm assemblies showed 8~9% higher than HiFi-PromethION hybrid assembly on haploid completeness. We stated this on the discussion section, line 325.<br><br>3. Could the authors use their dipcall analysis to better understand what is missing from the assembly (e.g., segmental duplications)?<br><br>To identify missing regions, we made an alignment of our assembly against CHM13 v1.1 using Mummer and Dot. From an alignment against CHM13, we found long missing sequences on centromeric regions and they could be found on Fig. S1 (chr. 1) and S2 (chr. X).<br><br>4. I suspect the assembly may collapse many segmental duplications, causing base-level and structural errors in the assembly, which could cause many problems when using the reference, so the authors should make this clear. For example, how many of the missing genes are in segmental duplications?<br><br>From an alignment against CHM13, we found long missing sequences on centromeric regions and they could be found on Fig. S1 (chr. 1) and S2 (chr. X). On chromosome one, about 29 Mb was missing and they were located on a centromeric region. On chromosome X, missing sequences of a centromeric region had a length of about 4 Mb (Fig. S2). We stated this on Line 242.<br><br>5. What version of hifiasm was used by the authors?<br><br>We used v0.15.5-r352. |

6. This statement is mis-leading, since the CHM13 reference is much more complete and contiguous, even though KOREF has a comparable scaffold length: "The results showed that KOREF_S1v2.1 is more contiguous than AK1 and HuRef, and comparable to JG2.0.0 Beta and CHM13_v1.1 (Table 2). Among six genome assemblies, KOREF_S1v2.1 and CHM13 were a haplotype-resolved assembly with a chromosome-scale". It should be revised to something like "The results showed that KOREF_S1v2.1 has longer scaffold N50 than AK1 and HuRef, and scaffold N50 comparable to JG2.0.0 Beta and CHM13_v1.1 (Table 2). Among these six genome assemblies, KOREF_S1v2.1 and CHM13 were the only haplotype-resolved assemblies with a chromosome-scale, though KOREF_S1v2.1 has lower QV, shorter contigs, and is missing 8-10% of the human genome sequence included in CHM13_v1.1. KOREF_S1v2.1 also has longer scaffolds than recent trio-hifiasm-based assemblies, but has shorter contig N50, lower QV, and substantially lower completeness."

We agree with your comments and revised the texts according to your suggestion.

7. Could the authors please elaborate on this conclusion "From a pilot study, an error-correction module of the 3D-DNA pipeline seemed to split long repetitive regions complicatedly, and it made difficult to construct scaffolds or curate misassembles (Fig. S1)"? No Fig S1 was included in the submission, and this merits more discussion and detailed methods if the authors want to claim this.

You are right. We missed to include Fig. S1 and prepared it as Fig. S3-A and -B. We constructed scaffolds using contigs from KOREF's paternal hifiasm assembly and Hi-C sequencing data by 3D-DNA pipeline. Fig. S3-A shows a Hi-C heat map of contigs without correcting misassemblies and Fig. S3-B shows a Hi-C heat map of contigs/scaffolds with correcting misassemblies. On Fig. S3-A, we can find white stripe patterns from long repetitive regions, such as centromeres or telomeres, in contigs or on the border of contigs. However, on Fig. S3-B, a small number of white stripes were found in scaffolds. And we found a large amount of short contigs with long repetitive sequences that have appeared to come from centromeres or telomeres. Its length reaches 160 Mb. The developers of 3D-DNA pipeline already have warned this on their github page. To avoid this problem, we needed to build a new strategy that enabled to correct local misassemblies on long repetitive regions by Hi-C sequencing. We noted this on Line 286.
8. The authors should state in the main text that the N50 read lengths from ONT and HiFi, since they are relatively small compared to current best practice.

Good point. We added N50 read lengths and the longest read lengths from ONT and HiFi in the results section (Line 180). An N50 of PromethION sequencing ranged from 6,793 bp to 18,109 bp and an N50 of PacBio HiFi ranged from 11,846 bp to 15,901. About lengths of the longest read, PromethION ranged from 160,294 bp to 1,753,381 bp and PacBio HiFi ranged from 28,947 bp to 36,401 bp.

9. It would be useful to compare to other recent reference genomes and assemblies, such as https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02047-7, https://doi.org/10.1101/2021.06.10.447952, and https://www.nature.com/articles/s41592-020-01056-5

Thank you for recommending additional human genome assemblies. We have now added some comparison statistics of Ash1 assembly and PR1 assembly to table 2 (Line 231). The results showed that KOREF_S1v2.1 has longer scaffold N50 than AK1, HuRef, Ash1 and PR1, and scaffold N50 was comparable to JG2.0.0 Beta and CHM13_v1.1.

Reviewer #2:
This paper reported the construction of KOREF_S1v2, a reference genome for Korean or Eastern Asian, using long read sequencing platforms in addition to NGS sequencing and HiC sequencing platforms. A reference genome construction method was introduced to combine parent genomes to increase the quality of the final assembled genome. The constructed genome was assessed for its quality by comparing it to the existing KOREF genome and the human reference genome. The goal of this paper is to provide an accurate Korean reference genome. A few issues are listed below.

| | |
|---|---|
| | Major issues<br>1. Line 52, "GRCh38 … derives from a single individual, mostly based on Caucasian and African ancestry", the content was incorrect, and the sentence needs a revision.<br><br>You are right. GRCh38 was constructed from thirteen anonymous volunteers. We corrected the Line 52 from “a single individual” to “thirteen anonymous volunteers’. Thank you.<br><br>2. Line 200, "the genes included 20,378 protein-coding genes with 166,570 transcripts, 46,973 lncRNAs and 17,535 pseudogenes.", the number of protein coding transcripts, 166,570, was much bigger than the number for protein-coding transcripts listed in GENCODE, which is about 87K. Please double check the numbers.<br><br>We agree with you. We found a mistake on the liftover and have fixed it. Now, we have 19,668 protein coding genes with 85,889 transcripts (Line 217). Thank you.<br><br>Minor issues<br>1. Line 234, "and it made difficult to construct" -->" and made it difficult to construct".<br><br>Thank you. We fixed the text as your suggestion. |
| **Additional Information:** | |
| **Question** | **Response** |
| Are you submitting this manuscript to a special series or article collection? | No |
| **Experimental design and statistics**<br><br>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.<br><br>Have you included all the information requested in your manuscript? | Yes |
| **Resources**<br><br>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. | Yes |

| | |
|---|---|
| Have you included the information requested as detailed in our Minimum Standards Reporting Checklist? | |
| **Availability of data and materials**<br><br>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.<br><br>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist? | Yes |

# KOREF_S1: the phased, parental Trio-binned Korean reference genome using long-reads and Hi-C sequencing methods

Hui-su Kim[1], Sungwon Jeon[1,2,3], Yeonkyung Kim[1,3], Changjae Kim[1,3], Jihun Bhak[1,2], and Jong Bhak[1,2,3,4]*

[1)] Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, Republic of Korea

[2)] Department of Biomedical Engineering, College of Information and Biotechnology, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, Republic of Korea

[3)] Clinomics Inc, Ulsan, 44919, Republic of Korea

[4)]Personal Genomics Institute, Genome Research Foundation, Cheongju, 28160, Republic of Korea

*Correspondence author:

Name: Jong Bhak, Ph.D.

Address: #110-303, Ulsan National Institute of Science and Technology, UNIST-gil 50, Eonyang-eup, Ulju-gu, Ulsan 44919, Republic of Korea

Phone: (+82) 10-4644-6754

Email: jongbhak@genomics.org

**ORCID iDs:**

Jong Bhak [0000-0002-4228-1299]; Hui-su Kim [0000-0003-2277-638X]; Sungwon Jeon [0000-0002-2729-9087];

# Abstract

## Background

KOREF is the Korean reference genome which was constructed with various sequencing technologies including long reads, short reads, and optical mapping methods. It is also the first East Asian multiomic reference genome accompanied by extensive clinical information, time series and multiomic data, and his parental sequencing data. However, it was still not a chromosome-scale reference. Here, we updated the previous KOREF assembly to a new chromosome-level haploid assembly of KOREF, KOREF_S1v2.1. ONT PromethION, PacBio HiFi-CCS, and Hi-C technology were used to build the most accurate East Asian reference assembled so far.

## Results

We produced 705 Gb ONT reads and 114 Gb PacBio HiFi reads, and corrected ONT reads by PacBio reads. The corrected ultra-long reads reached higher accuracy of 1.4% base-errors than the previous KOREF_S1v1.0, which was mainly built with short reads. KOREF has parental genome information, and we successfully phased it using a trio-binning method acquiring a near-complete haploid-assembly. The final assembly resulted in total length of 2.9 Gb with an N50 of 150 Mb,

and the longest scaffold covered 97.3% of GRCh38's chromosome 2. And the final assembly showed high base accuracy, less than 0.01% of base-errors.

## Conclusions

KOREF_S1v2.1 is the first chromosome-scale haploid assembly of the Korean reference genome with high contiguity and accuracy. Our study provides useful resources of the Korean reference genome and demonstrates a new strategy of hybrid assembly which collaborates ONT's PromethION and PacBio's HiFi-CCS.

**Keywords:** Korean reference; KOREF_S1; ONT PromethION; PacBio HiFi; Hi-C; hybrid assembly

# Introduction

Since the human genome reference was released in 2003, it has been updated and recently was patched in 2019 (GRCh38.p13) by the Genome Reference Consortium (GRC) [1]. Despite high completeness of GRCh38 assembly, it derives from thirteen anonymous volunteers, mostly based on Caucasian and African ancestry [2]. It is the most precise and extensive among all human references constructed so far. Recently, due to recent cost-effective sequencing methods, especially long reads methods, one can construct human personal references fast and efficiently [3]. The first Korean reference, KOREF, has been constructed in two types [4]. The first is KOREF_S1 which is a personal reference from an individual which is accompanied by parental *de novo* assemblies. The second one is KOREF_C which is a consensus population reference that includes variome information of Koreans. KOREF was initiated by the Korean Ministry of Science and Technology in 2006 to generate a national genome and variome references and currently it is jointly developed by the Genome Research Foundation, National Standard Reference Research Center, and the Korean Genomics Center at UNIST (Ulsan National Institute of Science and Technology). The first version of KOREF_S1, KOREF_S1v1.0, had a clear limitation of short reads and long-distance mapping-based approaches that resulted in a relatively low-quality assembly compared to the current GRCh38. We used Oxford Nanopore Technologies (ONT) PromethION and PacBio HiFi sequencers to upgrade KOREF_S1 by using a publicly available KOREF cell line.

# Materials and Methods

## Sample preparation and genome sequencing

Sample preparation steps were followed as the previous study [4, 5]. Human KOREF cell lines [6] were cultured at 37°C in 5% $CO_2$ in RPMI-1640 medium with 10% heat-inactivated fetal bovine serum. DNA was extracted from cells using the DNeasy Blood & Tissue kit (Qiagen) to the manufacturer's instructions. Sequencing libraries for the Oxford Nanopore Technologies PromethION were prepared using the 1D ligation sequencing kit (SQK-LSK109, Oxford Nanopore Technologies, UK) following the manufacturer's instructions. The products were quantified using the Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA) and raw signals were generated by the PromethION R9.4.5 platform (RRID:SCR_017987, Oxford Nanopore Technologies, UK). Base-calling the raw signals was performed using Guppy v4.0.11 with the Flip-flop hac model.

Genomic DNA from KOREF blood samples was extracted using QIAGEN Blood & Cell Culture DNA Kit (cat no 13323). A total of 5 μg of each sample was used as input for library preparation. The SMRTbell library was constructed using the SMRTbell® Express Template Preparation Kit (101-357-000). Using the BluePippin Size selection system we removed the small fragments for a large-insert library. After sequencing primer v4 was annealed to the SMRTbell template, DNA polymerase was bound to the complex (Sequel Binding kit 2.0). We purified the complex using AMPure Purification to remove excess primer and polymerase prior to sequencing. The SMRTbell library was sequenced using SMRT cells (Pacific Biosciences) using Sequel Sequencing Kit v2.1 and 10 hr movies were captured for each SMRT Cell 1M v2 using the Sequel II (RRID:SCR_017990, Pacific Biosciences) sequencing platform.

Hi-C libraries were generated using the Arima-Hic kit (A160105v01, San Diego, CA, USA). KOREF cell lines and blood samples were prepared for the construction of Hi-C libraries. Briefly,

chromatin from cross-linked cells was solubilized and then digested using restriction enzymes MboI or Arima's multiple enzymes (GATC and GANTC). The digested ends were labeled using a biotinylated nucleotide, and ends were ligated to create ligation products. Ligation products were purified, fragmented, and selected by size using AMPure XP Beads. Illumina-compatible sequencing libraries were constructed on end repair, dA-tailing, and adaptor ligation using a modified workflow of the Hyper Prep kit (KAPA Biosystems, Inc.). The bead-bound libraries were amplified and purified using AMPure XP beads and sequenced using Illumina NovaSeq platform with a read-length of 150 bp by Novogene (Beijing, China). Short paired-end raw reads using Illumina HiSeq 2000 platform were acquired from a previous study, accession no. SRR2204706. For generating parental sequencing reads, we prepared samples from both of KOREF_S1's parents. DNA was extracted from the donor's blood using DNAeasy Blood & Tissue Kit from QIAGEN according to the manufacturer's instruction. The quality and concentration of the extracted DNA were evaluated using NanoDrop™ One/OneC UV-Vis Spectrophotometer (Thermo Scientific™). Library construction and whole-genome sequencing were performed by Illumina HiSeq platform (Illumina, USA) with a 100 bp paired-end sequencing.

## Preprocessing of sequenced reads

The sequenced long- and short-read data were performed preprocessing steps as adapter trimming, quality trimming, and error correction. For the long reads, adapter trimming was performed using Porechop v0.2.4 (Porechop, RRID:SCR_016967) [7] and removing reads with below quality-score 7 was performed using Guppy. For the short reads, adapter- and quality trimming were performed using Trimmomatic v0.39 (Trimmomatic, RRID:SCR_011848) [8], and an error correction was

119 performed using the tadpole.sh program of BBtools suite v38.26 (Bestus Bioinformaticus Tools,
120 RRID:SCR_016968) [9].

121

122 ## Trio-binning and read correction

123 To obtain more accurate and longer haplotype-resolved reads from ONT PromethION sequencing,
124 we applied a trio-binning with KOREF's parental sequencing data and an error-correction with
125 PacBio HiFi sequencing data. The whole procedure is described in figure 1. To obtain haplotype-
126 resolved reads from ONT PromethION and PacBio HiFi sequencing, we performed a trio-binning
127 using TrioCanu v2.1 (Canu, RRID:SCR_015880) [10] with the parental short-reads. In this step,
128 reads from eleven PromethION flow-cells and six PacBio HiFi cells were participated. We merged
129 unclassified reads to the classified paternal-reads and classified maternal-reads each. To correct
130 base-errors on the PromethION reads, we corrected the errors with the haplotype-resolved reads
131 from PacBio HiFi sequencing using Racon v1.4.3 (Racon, RRID:SCR_017642) [11]. We acquired
132 KOREF's parental sequencing data from the KOREF homepage [6].

133

134 ## *De novo* assembly of KOREF_S1 genome

135 Contig assembly was processed using wtdbg2 v2.5 [12] (WTDBG, RRID:SCR_017225) and Flye
136 assembler v2.8.1 (Flye, RRID:SCR_017016) [13]. For a wtdbg2 assembly, parameters were set as
137 '-x corrected -g 3g -L 5000 -X 70.0'. An error correction of the assembled contigs was conducted
138 using Racon with a single iteration. The Flye assembly was performed with parameters of '--
139 pacbio-hifi --hifi-error 0.008 --genome-size 3g'. For error correction, we carried out the same
140 procedure as the wtdbg2 assembly.

To construct scaffolds with a chromosome-scale, we conduct scaffolding using PromethION reads and Hi-C data. To scaffold contigs using PromethION reads, LINKS v1.8.7 [14] was used with a single flow-cell of PromethION reads. To construct chromosome-scale scaffolds using Hi-C data. 3D-DNA pipeline v180922 [15] with Juicer v1.6.2 program (Juicer, RRID:SCR_017226) [16] was performed with the scaffolds by LINKS. Hi-C raw reads were mapped against the extended contigs using Juicer, and the 3D-DNA pipeline was initiated to correct mis-joined contigs and construct scaffolds. To correct misassemblies on the scaffolds, a manual curation was performed using JBAT (JuiceBox Assembly Tool) v1.11.08 program (Juicebox, RRID:SCR_021172) [17]. To polish base-errors and small indels, we performed Pilon v1.23 program (Pilon, RRID:SCR_014731) [18] with KOREF's short read data and parameters of '--fix snps and indels' were used.

## Constructing high-confident regions, and the assessment of base-errors on long-reads and genome assemblies

For an assessment of base-errors, we constructed high-confident regions of KOREF_S1 v1 against chromosome sequences of the GRCh38.p13. The procedure was referred to Heng Li's study [19]. We aligned the KOREF_S1v1.0 assembly to GRCh38 using the Minimap2 program v2.17-r941 (Minimap2, RRID:SCR_018550) [20]. Alignments with mapping quality >5 and aligned segments shorter than 50 kb were discarded. The filtered alignments were converted to the BED format and sorted.

To assess base-errors of long-reads and genome assemblies, we compared them to the KOREF_S1v1.0 assembly using the assembly_assess program from Pomoxis v0.3.4 [21]. And the Merqury v1.0 [22] program was performed to assess assemblies using k-mers.

## Genome annotation

To identify protein coding genes on KOREF_S1v2.1 genome, we performed a liftover with a gene annotation from GENCODE 38. The liftover was processed using Liftoff v1.6.1 program [23]. The result of genome annotation was stored in the KOREF genome browser, built by the JBrowse v1.16.9 (JBrowse, RRID:SCR_001004) [24]. To assess protein-coding genes, BUSCO analysis (BUSCO, RRID:SCR_015008) [25] was performed using BUSCO v5.2.2 and mammalian orthoDB v10.

# Results

## KOREF_S1v2.1 assembly

We obtained 235× coverage (705 Gb) of long-reads from twelve ONT PromethION flow-cells and 38× coverage (114 Gb) of long reads from six PacBio HiFi cells (Table S1). We also acquired 274 Gb corrected paternal haplotype-resolved reads and 265 Gb corrected maternal haplotype-resolved reads after trio-binning and read-correction. An N50 of PromethION sequencing ranged from 6,793 bp to 18,109 bp and an N50 of PacBio HiFi ranged from 11,846 bp to 15,901 bp. About lengths of the longest read, PromethION ranged from 160,294 bp to 1,753,381 bp and PacBio HiFi ranged from 28,947 bp to 36,401 bp. The corrected reads were identified about 1.4% base-errors (Table S2). Contigs from both haplotypes were assembled using wtdbg2 and Flye. The Flye assembly showed better results of higher N50 values (19.47 Mb for a paternal and 25.86 Mb for a

maternal assembly) and longer length of the longest contig (70.97 Mb for a paternal and 109.79 Mb for a maternal assembly) (Table 1).

We extended the contigs to chromosome-scale scaffolds using 76.5 Gb of PromethION reads (Flow-cell no.2) and 884 Gb of Hi-C data (294× sequencing-depth). Scaffolds from a mitochondrial genome were excluded using the KOREF's mtDNA sequence from the previous study [4]. As a result, we acquired the paternal assembly of 2.82 Gb length with 2,230 scaffolds and an N50 of 141.04 Mb (Table 1). The maternal assembly resulted in 2,616 scaffolds with an N50 of 150.05 Mb, and its total length was 2.88 Gb. For generating the final assembly of KOREF_S1v2.1, we substituted sequences of autosomal chromosomes and a Y chromosome from the paternal assembly, and a X chromosome from the maternal assembly. As a result, the KOREF_S1v2.1 was acquired a total length of 2.9 Gb with an N50 of 150.05 Mb.

## Genome annotation

We annotated genes in KOREF_S1v2.1 by integrating a liftover of gene annotations from the GENCODE release 38 (https://www.gencodegenes.org/human/) and homology information of RNASeq data. The genes included 19,668 protein-coding genes with 85,889 transcripts, 46,973 lncRNAs and 17,535 pseudogenes (Table 3). From assessment of protein-coding genes by BUSCO, 99.3% of complete orthologous genes were found and 0.6% were missing (Table 4). 1,391 genes from the Gencode38 annotation were not transferred to the KOREF by liftover, and a list of these genes can be found in the supplementary table 4.

## Assessment of KOREF and comparison with other human genome assemblies

Using the Merqury program for a quality assessment, we estimated QV scores of Q43.88 for the paternal assembly and Q44.49 for the maternal assembly. The final assembly showed QV score of Q43.88, indicating >99.99% accuracy (Table S5), and it is higher than KOREF_S1v1.0's (Q33.58) and KOREF_S1v2.0 (Q39.52) which was were assembled with the PromethION data. We compared KOREF_S1v2.1 and other human reference genome assemblies (AK1_v2, JG2.0.0 Beta, HuRef, CHM13_v1.1, GRCh38.p13, Ash1v2.0 and PR1 v3.0) [26-31]. The results showed that KOREF_S1v2.1 has a longer scaffold N50 than AK1, HuRef, Ash1 and PR1, and scaffold N50 was comparable to JG2.0.0 Beta and CHM13_v1.1 (Table 2). Among these eight genome assemblies, KOREF_S1v2.1 and CHM13 were the only haplotype-resolved assemblies at a chromosome-scale, though KOREF_S1v2.1 has lower QV, shorter contigs, and is missing 8-10% of the human genome sequence included in CHM13_v1.1. KOREF_S1v2.1 also has longer scaffolds than recent trio-hifiasm-based assemblies, but has shorter contig N50, lower QV, and substantially lower completeness. AK1 was haplotype-resolved using a read-based phasing method but could not reach a chromosome-scale without a guidance of the reference genome.

To identify missing regions on KOREF_S1v2.1, we made an alignment plot of KOREF against CHM13 v1.1 using Mummer v4.0.0beta2 (MUMmer, RRID:SCR_018171) [32] and Dot [33]. We found long missing sequences on centromeric regions (Fig. S1). On chromosome one, about 29 Mb was missing and they were located on a centromeric region. On chromosome X, missing sequences of a centromeric region were a length of about 4 Mb (Fig. S2).

From a pilot study of KOREF_S1's PacBio HiFi sequencing by Hifiasm v0.15.5-r352 (Hifiasm, RRID:SCR_021069) [34], a contig assembly (KOREF_S1v2.0_PBCCS hifiasm_trio) resulted in

highest base-accuracy and contiguity between HiFi-only, PromethION, and HiFi-PromethION hybrid assembly (Table S3). About haploid completeness, it scored 99.6873% (maternal) and 99.1902% (paternal), which showed 8~9% higher than KOREF_S1v2.1 assembly.

For comparing assembly quality of HiFi, PromethION and HiFi-ProemthION hybrid, we compared contigs assemblies from HG00744, HG002, and KOREF. HiFi assemblies showed highest QV and NG50 (Table 5). HG002 assembly showed highest QV of 51.6 and PromethION assembly of KOREF showed lowest QV of 33.8. HiFi-PromethION hybrid assembly of KOREF scored higher QV (42.2) against PromethION assembly. But it was lower than the HiFi assembly of KOREF (QV 45.1).

# Discussion

In previous version of KOREF_S1, we generated a chromosome-level genome assembly with a guidance of GRCh38. A new version of KOREF assembly, KOREF_S1v2.1, was assembled with high accurate (less than 0.01% of base error) and contiguity from multiple sequencing technologies including ONT, PacBio, Illumina, and Hi-C. Furthermore, the new KOREF assembly was phased with parental sequencing data. To generate ultra-long and high accurate reads, we corrected ONT reads using PacBio HiFi reads. Most genomic regions were covered by the corrected reads, but some highly competitive regions including telomere and centromere were not covered. They were remained as gaps with unknown length. Especially on a chromosome Y, we found more gaps and less contiguity than other chromosomes. The genomic sequences of a chromosome X and Y have high similar regions and they probably make difficulties to phase genomic sequences on sex chromosomes.

Recently, new *de novo* assembly pipelines, such as the Hifiasm [34] and HiCanu [35], have been developed for PacBio's HiFi-CCS. Hifiasm supports a trio-binning from parental sequencing and Hi-C. From a pilot study by Hifiasm, a contig assembly of hifiasm_trio showed the highest base-accuracy and contiguity (Table S3). About haploid completeness, it also showed the highest value, 8~9% more against KOREF_S1v2.1. Despite these advantages, scaffolding contigs from Hifiasm has difficulties for using Hi-C data. Error-correction modules of the 3D-DNA pipeline seemed to split long repetitive sequences complicatedly and made it difficult to construct scaffolds or curate misassembles (Fig. S3). Fig. S3-A shows a Hi-C heat map of contigs without correcting misassemblies and Fig. S3-B shows a Hi-C heat map of contigs/scaffolds with correcting misassemblies. On Fig. S3-A, we can find white stripe patterns from long repetitive regions, such as centromeres or telomeres, in contigs or on the border of contigs. However, on Fig. S3-B, a small number of white stripes were found in scaffolds. And we found a large amount of short contigs with long repetitive sequences that have appeared to come from centromeres or telomeres. Its length reaches 160 Mb. The developers of 3D-DNA pipeline already have warned this on their github page. In order to avoid this problem, we needed a new strategy that enabled to correct local miassemblies on long repetitive regions by Hi-C sequencing. However, the high-quality contigs from Hifiasm can be helpful to remove gaps and showed possibility to resolve highly repetitive regions. Also, a recent study of the T2T consortium shared a complete structure of centromeric regions [29], and it will be a useful resource to complete the KOREF_S1 genome.

In conclusion, we upgraded a high-quality Korean reference genome, KOREF. Our study provides useful resources of the Korean reference genome and demonstrates a new strategy of hybrid assembly which collaborates ONT's PromethION and PacBio's HiFi-CCS.

## Data availability

The Korean reference genome project has been deposited at DDBJ/ENA/GenBank under the accession PRJNA735947. The version described in this paper is version JAHRJT000000000. Raw DNA and RNA sequence reads for KOREF and KPGP have been submitted to the NCBI Sequence Read Archive database (Table S1). The immortalized cell line of KOREF was deposited in the Korean Cell Line Bank (KCLB, #60211). KOREF_S1 data is found from the Korean Reference Genome Project website [36]. All supporting data and materials are available in the *GigaScience* GigaDB database [37].

## Abbreviations

KOREF: KOrean REFerence

ONT: Oxford Nanopore Technologies

BUSCO: Benchmarking Universal Single-Copy Orthologs

## Competing financial interests

The authors declare no competing financial interests.

## Funding

This work was supported by the Promotion of Innovative Businesses for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (1425157253) (2.220037.01). This work was also supported by the Establishment of Demonstration Infrastructure for

Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (1425157301) (2.220036.01).

## Author contributions

J.B. supervised and coordinated the national Korean reference genome project and Personal Genome Project Korea. J.B. conceived and designed the reference genome project. H.K. performed the analyses and assembly. H.K. and J.B. wrote the manuscript.

## 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). 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. We thank Jaesu Bhak for editing the manuscript.

## References

1. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27 5:849-64. doi:10.1101/gr.213611.116.
2. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328 5979:710-22. doi:10.1126/science.1188021.
3. Logsdon GA, Vollger MR and Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21 10:597-614. doi:10.1038/s41576-020-0236-x.
4. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun. 2016;7 doi:ARTN 13637 10.1038/ncomms13637.
5. Kim H, Jeon S, Kim C, Kim Y, Cho YS, Kim J, et al. Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information. GigaScience. 2019;8:12 doi:10.1093/gigascience/giz125.
6. KOREF_S Resource homepage. http://koref.net.
7. Ryan W, Jeremy V. Porechop (Version 0.2.4). 2018. https://github.com/rrwick/Porechop.
8. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30 15:2114-20. doi:10.1093/bioinformatics/btu170.
9. JGI. BBTools (Version 38.26). 2018. https://jgi.doe.gov/data-and-tools/bbtools/.
10. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018; doi:10.1038/nbt.4277.
11. Vaser R, Sovic I, Nagarajan N and Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27 5:737-46. doi:10.1101/gr.214270.116.
12. Ruan J and Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17 2:155-+. doi:10.1038/s41592-019-0669-3.
13. Kolmogorov M, Yuan J, Lin Y and Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37 5:540-+. doi:10.1038/s41587-019-0072-8.
14. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience. 2015;4 doi:ARTN 35 10.1186/s13742-015-0076-3.
15. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356 6333:92-5. doi:10.1126/science.aal3327.
16. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016;3 1:95-8. doi:10.1016/j.cels.2016.07.002.
17. Aidenlab. Juicebox (Version 1.11.08). 2019. https://github.com/aidenlab/Juicebox.

18. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Plos One. 2014;9 11 doi:ARTN e112963 10.1371/journal.pone.0112963.
19. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat Methods. 2018;15 8:595-7. doi:10.1038/s41592-018-0054-7.
20. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34 18:3094-100. doi:10.1093/bioinformatics/bty191.
21. Oxford Nanopore. Pomoxis (Version 0.3.4). 2020. https://github.com/nanoporetech/pomoxis.
22. Rhie A, Walenz BP, Koren S and Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21 1:245. doi:10.1186/s13059-020-02134-9.
23. Shumate A and Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2020; doi:10.1093/bioinformatics/btaa1016.
24. Robert B, Eric Y, Colin MD, Richard DH, Monica MT, Gregg H, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17,66. doi:10.1186/s13059-016-0924-1.
25. Mosè M, Matthew RB, Mathieu S, Felipe AS, Evgeny MZ. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. MBE. 2021;38 10:4647-54. doi:10.1093/molbev/msab199.
26. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016;538 7624:243-+. doi:10.1038/nature20098.
27. Takayama J, Tadaka S, Yano K, Katsuoka F, Gocho C, Funayama T, et al. Construction and integration of three de novo Japanese human genome assemblies toward a population-specific reference. Nat Commun. 2021;12 1:226. doi:10.1038/s41467-020-20146-8.
28. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5 10:e254. doi:10.1371/journal.pbio.0050254.
29. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. bioRxiv. 2021.
30. Alaina S, Aleksey VZ, Rachel MS, Daniela P, Justin MW, Nathan DO, et al. Assembly and annotation of an Ashkenazi human reference genome. Genome Biol. 202;21:129. doi:10.1186/s13059-020-02047-7.
31. Aleksey Z, Alaina S, Ida S, Jakob H, Daniela P, Mihaela P, et al. A reference-quality, fully annotated genome from a Puerto Rican individual. Genetics. 2021;220:2. doi:10.1101/2021.06.10.447952.
32. Mummer4. Mummer (Version 4.0.0beta2). 2017. https://github.com/mummer4/mummer.
33. Maria N. Dot. 2020. https://github.com/marianattestad/dot.

34. Cheng H, Concepcion GT, Feng X, Zhang H and Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18 2:170-5. doi:10.1038/s41592-020-01056-5.
35. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30 9:1291-305. doi:10.1101/gr.263566.120.
36. The Korean Reference Genome (KOREF) Project homepage. 2006. http://koreanreference.org.
37. Kim H; Jeon S; Kim Y; Kim C; Bhak J; Bhak J. Supporting data for "KOREF_S1: the phased, parental Trio-binned Korean reference genome using long-reads and Hi-C sequencing methods" GigaScience Database 2022. http://doi.org/10.5524/100983.

# Figures

Figure 1. The flowchart of KOREF reference genome assembly.

Figure S1. Alignment of chr1 sequence, KOREF_S1v2.1 versus CHM13 v1.1.

Figure S2. Alignment of chr2 sequence, KOREF_S1v2.1 versus CHM13 v1.1.

Figure S3. Comparison of Hi-C heat map with and without correcting miassemblies by 3D-DNA pipeline.

A) A Hi-C heat map of KOREF_S1v2.0_PBCCS contigs (hifiasm_trio, paternal) by 3D-DNA pipeline without correcting misassemblies. Contigs were denoted by green boxes and scaffolds (or chromosomes) were denoted by blue boxes. Some contigs had white sprites and this means contigs with highly repetitive sequences such as centromeres or telomeres.

B) A Hi-C heat map of KOREF_S1v2.0_PBCCS contigs (hifiasm_trio, paternal) by 3D-DNA pipeline with correcting mis-assemblies. A small number of contigs/scaffolds had white sprites (long repetitive sequences). Instead, we found a large amount of short contigs with long repetitive sequences that have appeared to come from centromeres or telomeres (black box).

# Tables

**Table 1. The statistics of KOREF_S1v2.1 assembly.**

| | Contig | | | | Scaffold | |
|---|---|---|---|---|---|---|
| | **Wtdbg2_paternal** | **Flye_paternal** | **Wtdbg2_maternal** | **Flye_maternal** | **Paternal** | **Maternal** |
| Sequence no. | 3,059 | 2,973 | 2,426 | 2,475 | 2,230 | 2,616 |
| Total length (bp) | 2,652,350,533 | 2,820,210,305 | 2,691,371,348 | 2,885,670,065 | 2,821,407,033 | 2,886,600,011 |
| N50 (bp) | 15,085,508 | 19,472,363 | 15,312,743 | 25,861,606 | 141,044,433 | 150,051,441 |
| Longest (bp) | 70,969,653 | 87,371,841 | 70,444,093 | 109,786,075 | 235,665,501 | 234,237,609 |
| Gaps | 0.000% | 0.000% | 0.000% | 0.000% | 0.048% | 0.037% |
| GC contents | 40.90% | 40.92% | 40.84% | 40.86% | 40.92% | 40.88% |

**Table 2. Comparison between KOREF and other human genomes.**

| | KOREF_S1v2.1 | AK1_v2 | JG2.0.0 Beta | HuRef | CHM13 v1.1 | GRCh38.p13 | Ash1v2.0 | PR1 v3.0 |
|---|---|---|---|---|---|---|---|---|
| Scaffolds no. | 2,230 | 2,832 | 1,173 | 4,530 | 24 | 472 | 334 | 89 |
| Total length (bp) | 2,901,828,151 | 2,904,207,228 | 3,059,652,438 | 2,844,000,504 | 3,054,832,041 | 3,272,089,205 | 3,188,555,634 | 3,116,169,811 |
| Scaffold N50 (bp) | 150,051,441 | 44,846,623 | 152,668,378 | 143,733,266 | 154,259,566 | 67,794,783 | 146,254,838 | 149,697,505 |
| Phasing approach | De novo | De novo | De novo | Reference-guided | De novo | De novo | Reference-guided | De novo* |
| Assembly level | Chromosome | Scaffold | Chromosome | Chromosome | Chromosome | Chromosome | Chromosome | Chromosome |
| Haplotype-resolved | Trio-binning | Read-based | No | No | Haploid cell line | No | No | No |

* PR1 v3.0 assembly used CHM13 assembly as a reference genome to remove gaps.

**Table 3. The statistics of KOREF reference genome annotation.**

| | KOREF_S1v2.1 gene |
|---|---|
| Genes no. | 19,668 |
| Transcripts no. | 85,889 |
| Total length of transcripts (bp) | 110,601,598 |
| N50 (bp) | 1,983 |
| Length of longest transcripts | 107,976 |
| GC contents | 51.60% |
| lncRNAs no. | 46,973 |
| Pseudogenes no. | 17,535 |

**Table 4. Statistics of KOREF_S1v2.1 protein coding genes using BUSCO.**

| BUSCO assessment | KOREF_S1v2.1 protein coding genes |
|---|---|
| Complete | 99.3% |
| Complete and single-copy | 40.9% |
| Complete and duplicated | 58.4% |
| Fragmented | 0.1% |
| Missing | 0.6% |

**Table 5. Comparison of contigs from HG00733, HG002 and KOREF assembly.**

| Dataset | Seq. platform | Assembly | Size (Gb) | QV | NG50 (Mb) |
|---|---|---|---|---|---|
| HG00733 | PB HiFi | Hifiasm (trio) | 6.071 | 49.9 | 34.9 |
| HG002 | PB HiFi | Hifiasm (trio) | 5.967 | 51.6 | 43.0 |
| KOREF | PB HiFi | Hifiasm (trio) | 5.927 | 45.1 | 55.4 |
| KOREF | PromethION R9.4.1 | wtdbg2 (trio) | 5.527 | 33.8 | 9.3 |
| KOREF | PB HiFi - PromethION hybrid | Flye (trio) | 5.706 | 42.2 | 16.5 |

**Table S1. The statistics of sequencing data for KOREF assembly**

| Library name | Library type | Sequencer | No of reads | Total length of reads |
|---|---|---|---|---|
| KOREF_FC0 | Long read | ONT PromethION | 3,312,776 | 20,865,481,179 |
| KOREF_FC1 | Long read | ONT PromethION | 5,042,850 | 41,682,677,298 |
| KOREF_FC2 | Long read | ONT PromethION | 12,338,494 | 76,503,664,100 |
| KOREF_FC3 | Long read | ONT PromethION | 10,235,778 | 45,485,905,527 |
| KOREF_FC4 | Long read | ONT PromethION | 16,631,685 | 61,519,105,457 |
| KOREF_FC5 | Long read | ONT PromethION | 18,646,872 | 94,424,194,828 |
| KOREF_FC6 | Long read | ONT PromethION | 17,505,287 | 107,134,750,131 |
| KOREF_FC7 | Long read | ONT PromethION | 15,334,968 | 56,295,215,337 |
| KOREF_FC8 | Long read | ONT PromethION | 2,355,945 | 11,961,112,778 |
| KOREF_FC9 | Long read | ONT PromethION | 12,904,534 | 87,895,961,842 |
| KOREF_FC10 | Long read | ONT PromethION | 8,241,801 | 45,637,914,742 |
| KOREF_FC11 | Long read | ONT PromethION | 9,237,075 | 55,586,698,463 |
| KOREF_PBCCS_FC1 | Long read | PacBio HiFi-CCS | 851,009 | 13,490,105,391 |
| KOREF_PBCCS_FC2 | Long read | PacBio HiFi-CCS | 716,451 | 11,376,688,499 |
| KOREF_PBCCS_FC3 | Long read | PacBio HiFi-CCS | 951,653 | 15,081,743,678 |
| KOREF_PBCCS_FC4 | Long read | PacBio HiFi-CCS | 2,250,696 | 31,928,570,225 |
| KOREF_PBCCS_FC5 | Long read | PacBio HiFi-CCS | 1,637,317 | 19,173,274,135 |
| KOREF_PBCCS_FC6 | Long read | PacBio HiFi-CCS | 1,966,687 | 23,104,807,610 |
| K5_mbol | Hi-C | Illumina | 678,687,678 | 101,803,151,700 |
| WBC_mbol | Hi-C | Illumina | 1,030,159,562 | 152,405,701,620 |
| K5_combo | Hi-C | Illumina | 1,159,354,242 | 173,903,136,300 |
| WBC_combo | Hi-C | Illumina | 640,984,808 | 96,147,721,200 |
| KOREF_HiC | Hi-C | Illumina | 2,383,095,442 | 359,847,411,742 |

| N50 of reads | Length of the longest reads |
|---|---|
| 13,367 | 1,042,663 |
| 18,109 | 1,753,381 |
| 13,471 | 864,374 |
| 8,575 | 837,815 |
| 7,219 | 694,220 |
| 10,637 | 890,851 |
| 13,118 | 697,847 |
| 6,793 | 504,030 |
| 9,794 | 139,596 |
| 14,725 | 129,808 |
| 12,625 | 176,495 |
| 13,369 | 160,294 |
| 15,870 | 33,546 |
| 15,901 | 30,931 |
| 15,867 | 36,401 |
| 14,338 | 33,979 |
| 11,816 | 28,947 |
| 11,840 | 33,434 |
| 150 | 150 |
| 147 | 147 |
| 150 | 150 |
| 150 | 150 |
| 151 | 151 |

**Table S2. Base accuracy of raw and corrected ONT long-reads**

| | Error rate on Paternal | Error rate on Maternal |
|---|---|---|
| FC0 | 1.497% | 1.534% |
| FC1 | 1.539% | 1.537% |
| FC2 | 1.401% | 1.412% |
| FC3 | 1.206% | 1.200% |
| FC4 | 1.248% | 1.269% |
| FC5 | 1.297% | 1.288% |
| FC6 | 1.474% | 1.488% |
| FC7 | 1.278% | 1.294% |
| FC8 | 1.684% | 1.719% |
| FC9 | 1.370% | 1.370% |
| FC10 | 1.330% | 1.318% |
| FC11 | 1.509% | 1.533% |
| **Total** | 1.408% | |

**Table S3. Assessment of KOREF genome assemblies using Merqury**

| Assembly level | Assembly name | Haploid info. | QV | Error rate | Total completeness (%) | Haploid completeness (%) |
|---|---|---|---|---|---|---|
| Contigs | KOREF_S1v2.0_PT | Maternal | 33.8043 | 0.00041646 | 95.3942 | 88.3218 |
| | | Paternal | 33.7619 | 0.00042054 | 94.1992 | 89.4112 |
| | | Both | 33.7833 | 0.00041848 | 98.5486 | N/A |
| | KOREF_S1v2.0_PBCCS (wtdbg2) | Maternal | 44.0176 | 0.00003965 | 95.3942 | 88.3218 |
| | | Paternal | 43.9261 | 0.00004049 | 94.1992 | 89.4112 |
| | | Both | 43.9719 | 0.00004007 | 98.5486 | N/A |
| | KOREF_S1v2.0_PBCCS (hifiasm_trio) | Maternal | 45.1250 | 0.00003073 | 97.8435 | 99.6873 |
| | | Paternal | 45.1054 | 0.00003086 | 93.825 | 99.1902 |
| | | Both | 45.1154 | 0.00003079 | 99.8751 | N/A |
| | KOREF_S1v2.1 | Maternal | 43.7455 | 0.00004221 | 97.4445 | 90.5411 |
| | | Paternal | 41.0087 | 0.00007927 | 95.561 | 91.6752 |
| | | Both | 42.1805 | 0.00006053 | 99.3284 | N/A |
| Chromosomes | KOREF_S1v1.0 | Diploid | 33.5807 | 0.00043846 | 97.0471 | N/A |
| | KOREF_S1v2.0_PT | Maternal | 39.5177 | 0.00011175 | 96.9514 | 82.8504 |
| | | Paternal | 39.5263 | 0.00011153 | 95.0758 | 85.9729 |
| | | Both | 39.5219 | 0.00011164 | 98.9057 | N/A |
| | KOREF_S1v2.1 | Maternal | 44.4916 | 0.00003555 | 97.5563 | 90.5983 |
| | | Paternal | 43.3409 | 0.00004633 | 95.7252 | 91.8444 |
| | | Both | 43.8849 | 0.00004088 | 99.4084 | N/A |

**Table S4. The list of genes which were not partcipated in LiftOver**

| Gene name |
|---|
| ENSG00000225972.1 |
| ENSG00000251823.2 |
| ENSG00000263526.1 |
| ENSG00000278791.1 |
| ENSG00000281133.1 |
| ENSG00000281825.1 |
| ENSG00000264603.1 |
| ENSG00000255972.1 |
| ENSG00000256353.1 |
| ENSG00000258173.1 |
| ENSG00000277052.1 |
| ENSG00000238500.1 |
| ENSG00000282993.1 |
| ENSG00000211826.1 |
| ENSG00000211842.1 |
| ENSG00000211865.1 |
| ENSG00000211900.2 |
| ENSG00000211904.2 |
| ENSG00000211905.1 |
| ENSG00000211907.1 |
| ENSG00000211909.1 |
| ENSG00000211911.1 |
| ENSG00000211912.1 |
| ENSG00000211914.1 |
| ENSG00000211915.1 |
| ENSG00000211917.1 |
| ENSG00000211918.1 |
| ENSG00000211920.1 |
| ENSG00000211921.1 |
| ENSG00000211923.1 |
| ENSG00000211924.1 |
| ENSG00000211925.1 |
| ENSG00000211928.1 |
| ENSG00000211930.1 |
| ENSG00000211931.1 |
| ENSG00000211933.2 |
| ENSG00000223997.1 |
| ENSG00000225825.1 |
| ENSG00000227108.1 |
| ENSG00000227196.1 |
| ENSG00000227335.1 |

ENSG00000227800.1
ENSG00000228131.1
ENSG00000228985.1
ENSG00000232543.2
ENSG00000233655.1
ENSG00000236170.1
ENSG00000236597.1
ENSG00000237020.1
ENSG00000237197.1
ENSG00000237235.2
ENSG00000237547.1
ENSG00000240041.1
ENSG00000242472.1
ENSG00000242887.1
ENSG00000253808.1
ENSG00000253820.1
ENSG00000254045.1
ENSG00000257825.1
ENSG00000259016.1
ENSG00000270705.1
ENSG00000280494.2
ENSG00000221641.1
ENSG00000259302.1
ENSG00000259646.1
ENSG00000270185.1
ENSG00000270451.1
ENSG00000270824.1
ENSG00000270961.1
ENSG00000271317.1
ENSG00000271336.1
ENSG00000282089.1
ENSG00000282268.1
ENSG00000282520.1
ENSG00000282599.1
ENSG00000283888.1
ENSG00000264399.1
ENSG00000265561.2
ENSG00000266416.1
ENSG00000212051.1
ENSG00000284239.1
ENSG00000252191.1
ENSG00000273837.1
ENSG00000275726.1

ENSG00000282732.1
ENSG00000283673.1
ENSG00000211593.2
ENSG00000211594.2
ENSG00000211595.2
ENSG00000211596.3
ENSG00000211597.2
ENSG00000264041.2
ENSG00000277842.1
ENSG00000283591.1
ENSG00000271523.1
ENSG00000284125.1
ENSG00000211680.2
ENSG00000211684.2
ENSG00000238584.1
ENSG00000232167.1
ENSG00000239255.1
ENSG00000265483.1
ENSG00000249472.1
ENSG00000251816.1
ENSG00000266270.1
ENSG00000271544.1
ENSG00000264233.1
ENSG00000280665.1
ENSG00000211764.1
ENSG00000211765.1
ENSG00000211766.1
ENSG00000252866.1
ENSG00000282320.1
ENSG00000282420.1
ENSG00000282431.1
ENSG00000282780.1
ENSG00000284261.1
ENSG00000252521.1
ENSG00000276277.1
ENSG00000283146.1
ENSG00000221081.1
ENSG00000224931.4
ENSG00000272681.2
ENSG00000273773.1
ENSG00000275110.1
ENSG00000278803.2
ENSG00000279245.1

ENSG00000280249.1
ENSG00000275882.1
ENSG00000227232.5
ENSG00000278267.1
ENSG00000268020.3
ENSG00000236601.2
ENSG00000235146.2
ENSG00000237973.1
ENSG00000229344.1
ENSG00000240409.1
ENSG00000248527.1
ENSG00000198744.5
ENSG00000116721.9
ENSG00000231103.2
ENSG00000204510.5
ENSG00000207434.1
ENSG00000234064.1
ENSG00000232423.6
ENSG00000229571.7
ENSG00000179412.11
ENSG00000279169.3
ENSG00000237700.2
ENSG00000204480.8
ENSG00000186301.8
ENSG00000117122.14
ENSG00000224183.1
ENSG00000227207.2
ENSG00000237763.10
ENSG00000234441.1
ENSG00000174876.17
ENSG00000227408.1
ENSG00000238122.1
ENSG00000274642.1
ENSG00000277313.1
ENSG00000270392.2
ENSG00000273694.1
ENSG00000275933.1
ENSG00000286106.1
ENSG00000263353.3
ENSG00000198019.13
ENSG00000188610.12
ENSG00000276118.1
ENSG00000277095.1

ENSG00000252830.2
ENSG00000274927.1
ENSG00000233430.3
ENSG00000237503.2
ENSG00000234571.2
ENSG00000268074.1
ENSG00000275075.1
ENSG00000275767.1
ENSG00000232527.8
ENSG00000276442.1
ENSG00000276756.4
ENSG00000271567.1
ENSG00000227212.3
ENSG00000229002.1
ENSG00000279782.2
ENSG00000222854.1
ENSG00000273825.1
ENSG00000271644.1
ENSG00000215784.6
ENSG00000223612.3
ENSG00000256374.2
ENSG00000276216.1
ENSG00000201789.1
ENSG00000270339.3
ENSG00000274428.1
ENSG00000274408.1
ENSG00000244371.2
ENSG00000229828.2
ENSG00000231551.8
ENSG00000201183.1
ENSG00000285062.1
ENSG00000252515.2
ENSG00000177144.8
ENSG00000252656.1
ENSG00000201699.1
ENSG00000284842.1
ENSG00000226500.2
ENSG00000150337.14
ENSG00000233030.2
ENSG00000244057.5
ENSG00000187238.5
ENSG00000225217.1
ENSG00000236439.4

ENSG00000225483.1
ENSG00000196550.11
ENSG00000229509.1
ENSG00000272055.1
ENSG00000226113.1
ENSG00000234941.1
ENSG00000203496.9
ENSG00000228702.1
ENSG00000239152.1
ENSG00000273225.4
ENSG00000215097.3
ENSG00000204177.10
ENSG00000189090.8
ENSG00000264404.3
ENSG00000276544.1
ENSG00000226964.1
ENSG00000264717.5
ENSG00000252149.1
ENSG00000270025.2
ENSG00000278561.1
ENSG00000170324.21
ENSG00000230166.1
ENSG00000288603.1
ENSG00000222108.1
ENSG00000233197.1
ENSG00000271848.2
ENSG00000278616.1
ENSG00000273946.1
ENSG00000228055.3
ENSG00000213147.3
ENSG00000254468.2
ENSG00000232390.3
ENSG00000273813.1
ENSG00000284018.1
ENSG00000284306.1
ENSG00000283873.1
ENSG00000284546.1
ENSG00000244398.1
ENSG00000150244.12
ENSG00000285537.1
ENSG00000214414.9
ENSG00000168930.13
ENSG00000166013.11

ENSG00000254818.1
ENSG00000254655.1
ENSG00000204450.8
ENSG00000237706.4
ENSG00000204397.9
ENSG00000249054.2
ENSG00000171847.11
ENSG00000164845.16
ENSG00000214487.3
ENSG00000214826.5
ENSG00000212432.1
ENSG00000013573.17
ENSG00000257005.1
ENSG00000280208.1
ENSG00000279124.1
ENSG00000279730.2
ENSG00000279231.1
ENSG00000268486.5
ENSG00000276183.1
ENSG00000215604.3
ENSG00000233905.1
ENSG00000285576.1
ENSG00000227151.4
ENSG00000234278.3
ENSG00000283371.1
ENSG00000258233.1
ENSG00000257644.1
ENSG00000257175.2
ENSG00000257731.2
ENSG00000258076.1
ENSG00000257635.2
ENSG00000277529.1
ENSG00000278143.1
ENSG00000222036.8
ENSG00000225210.10
ENSG00000286614.1
ENSG00000278594.1
ENSG00000196143.4
ENSG00000277156.1
ENSG00000244306.11
ENSG00000287515.1
ENSG00000274827.4
ENSG00000278301.1

ENSG00000187537.13
ENSG00000274649.1
ENSG00000278184.1
ENSG00000257432.1
ENSG00000258324.2
ENSG00000257493.1
ENSG00000259045.1
ENSG00000129515.20
ENSG00000184227.8
ENSG00000258408.1
ENSG00000258605.1
ENSG00000226777.7
ENSG00000211966.2
ENSG00000211974.3
ENSG00000281990.1
ENSG00000283571.1
ENSG00000280411.1
ENSG00000259769.1
ENSG00000181984.11
ENSG00000278497.1
ENSG00000258916.2
ENSG00000258707.2
ENSG00000270685.1
ENSG00000258590.5
ENSG00000238478.1
ENSG00000258420.1
ENSG00000230031.10
ENSG00000270831.1
ENSG00000266545.1
ENSG00000258494.1
ENSG00000259698.1
ENSG00000258771.1
ENSG00000243059.3
ENSG00000258684.2
ENSG00000247765.2
ENSG00000278522.5
ENSG00000281347.3
ENSG00000287345.1
ENSG00000284834.1
ENSG00000280709.2
ENSG00000284988.1
ENSG00000279639.2
ENSG00000259324.2

ENSG00000285116.1
ENSG00000182974.3
ENSG00000279408.3
ENSG00000280655.1
ENSG00000258585.2
ENSG00000264902.1
ENSG00000283524.2
ENSG00000285405.1
ENSG00000284500.3
ENSG00000285135.1
ENSG00000259435.3
ENSG00000260739.1
ENSG00000223877.4
ENSG00000271288.1
ENSG00000275363.1
ENSG00000277865.4
ENSG00000277561.5
ENSG00000277515.1
ENSG00000283273.1
ENSG00000277867.1
ENSG00000277755.1
ENSG00000273976.2
ENSG00000277505.1
ENSG00000175676.15
ENSG00000273981.1
ENSG00000276941.1
ENSG00000273756.4
ENSG00000260399.1
ENSG00000261739.2
ENSG00000237850.7
ENSG00000261497.1
ENSG00000153684.15
ENSG00000276891.1
ENSG00000261524.1
ENSG00000276928.1
ENSG00000274532.1
ENSG00000260159.1
ENSG00000260053.2
ENSG00000183629.13
ENSG00000276955.1
ENSG00000261041.1
ENSG00000227717.4
ENSG00000188626.6

ENSG00000248334.6
ENSG00000260844.2
ENSG00000179938.12
ENSG00000273818.1
ENSG00000178081.12
ENSG00000207432.1
ENSG00000206972.1
ENSG00000186399.10
ENSG00000274424.1
ENSG00000178115.11
ENSG00000277031.1
ENSG00000259890.1
ENSG00000207430.1
ENSG00000201084.1
ENSG00000271078.1
ENSG00000260211.2
ENSG00000288627.1
ENSG00000261491.1
ENSG00000249931.4
ENSG00000275776.1
ENSG00000215304.3
ENSG00000261279.5
ENSG00000206987.1
ENSG00000261708.1
ENSG00000206127.11
ENSG00000274076.1
ENSG00000261375.1
ENSG00000254912.2
ENSG00000215252.12
ENSG00000237289.10
ENSG00000242866.10
ENSG00000166762.19
ENSG00000206991.1
ENSG00000259187.1
ENSG00000238845.1
ENSG00000212424.1
ENSG00000252117.1
ENSG00000278422.1
ENSG00000235370.6
ENSG00000272887.1
ENSG00000259538.1
ENSG00000197627.3
ENSG00000230373.8

ENSG00000259244.1
ENSG00000189136.9
ENSG00000277582.1
ENSG00000188388.10
ENSG00000183909.6
ENSG00000275771.1
ENSG00000248893.3
ENSG00000261523.1
ENSG00000254609.1
ENSG00000270734.1
ENSG00000263918.1
ENSG00000265537.1
ENSG00000207425.1
ENSG00000183426.17
ENSG00000275259.1
ENSG00000270580.5
ENSG00000277770.1
ENSG00000103226.19
ENSG00000263029.1
ENSG00000257381.3
ENSG00000265373.2
ENSG00000244257.5
ENSG00000183889.12
ENSG00000278221.1
ENSG00000277698.1
ENSG00000276484.1
ENSG00000285628.1
ENSG00000205746.9
ENSG00000257563.1
ENSG00000266454.1
ENSG00000185164.15
ENSG00000274025.1
ENSG00000277014.1
ENSG00000157106.17
ENSG00000183747.12
ENSG00000260201.2
ENSG00000260306.1
ENSG00000271623.1
ENSG00000205609.13
ENSG00000275429.1
ENSG00000278665.1
ENSG00000196502.12
ENSG00000181625.17

ENSG00000260280.5
ENSG00000213648.11
ENSG00000258150.6
ENSG00000257506.1
ENSG00000102879.16
ENSG00000278887.2
ENSG00000261444.1
ENSG00000260847.1
ENSG00000260649.1
ENSG00000260540.2
ENSG00000205456.11
ENSG00000183632.14
ENSG00000261391.1
ENSG00000260864.1
ENSG00000260845.1
ENSG00000261108.2
ENSG00000260644.6
ENSG00000260414.1
ENSG00000205457.11
ENSG00000260781.1
ENSG00000259842.2
ENSG00000284209.2
ENSG00000288300.1
ENSG00000221725.1
ENSG00000261816.1
ENSG00000260923.7
ENSG00000214946.14
ENSG00000179277.9
ENSG00000233090.1
ENSG00000227790.7
ENSG00000266302.6
ENSG00000276088.1
ENSG00000188933.16
ENSG00000266129.1
ENSG00000264892.1
ENSG00000226145.7
ENSG00000131885.17
ENSG00000273018.7
ENSG00000227689.1
ENSG00000154874.15
ENSG00000276532.1
ENSG00000230197.6
ENSG00000265746.1

ENSG00000262202.4
ENSG00000277947.1
ENSG00000281000.1
ENSG00000262319.1
ENSG00000197665.7
ENSG00000263934.5
ENSG00000189423.13
ENSG00000287490.1
ENSG00000230528.7
ENSG00000170298.16
ENSG00000229586.2
ENSG00000214822.8
ENSG00000231645.3
ENSG00000264422.1
ENSG00000214819.2
ENSG00000231258.2
ENSG00000264943.1
ENSG00000250462.8
ENSG00000261499.2
ENSG00000265889.1
ENSG00000277341.1
ENSG00000274487.2
ENSG00000274284.1
ENSG00000212659.1
ENSG00000229351.1
ENSG00000275616.1
ENSG00000278774.1
ENSG00000274862.1
ENSG00000274452.1
ENSG00000267198.1
ENSG00000176681.14
ENSG00000265411.1
ENSG00000260075.1
ENSG00000238083.8
ENSG00000087995.16
ENSG00000239246.3
ENSG00000136487.18
ENSG00000259533.2
ENSG00000204414.13
ENSG00000271974.1
ENSG00000232938.2
ENSG00000215512.9
ENSG00000222087.1

ENSG00000267541.1
ENSG00000282458.1
ENSG00000176695.8
ENSG00000281379.2
ENSG00000178464.6
ENSG00000267053.8
ENSG00000243130.8
ENSG00000204941.14
ENSG00000243137.8
ENSG00000183281.15
ENSG00000271097.1
ENSG00000273445.1
ENSG00000225933.1
ENSG00000234231.4
ENSG00000253278.1
ENSG00000241351.3
ENSG00000253497.1
ENSG00000253265.1
ENSG00000240382.3
ENSG00000254157.1
ENSG00000253732.1
ENSG00000211611.2
ENSG00000253578.1
ENSG00000253625.1
ENSG00000241294.1
ENSG00000244116.3
ENSG00000253998.3
ENSG00000243238.1
ENSG00000253860.1
ENSG00000253461.1
ENSG00000253487.1
ENSG00000253592.1
ENSG00000242371.1
ENSG00000283427.1
ENSG00000283196.2
ENSG00000251039.3
ENSG00000250036.1
ENSG00000239975.3
ENSG00000253191.1
ENSG00000253999.1
ENSG00000211623.2
ENSG00000254097.1
ENSG00000211625.2

ENSG00000241244.1
ENSG00000224041.3
ENSG00000278857.1
ENSG00000239819.2
ENSG00000143429.10
ENSG00000286156.1
ENSG00000237383.1
ENSG00000227265.2
ENSG00000213355.3
ENSG00000223564.1
ENSG00000229089.8
ENSG00000277701.5
ENSG00000240069.1
ENSG00000204588.5
ENSG00000235721.1
ENSG00000229730.2
ENSG00000283283.2
ENSG00000282785.1
ENSG00000263881.1
ENSG00000266139.2
ENSG00000223973.2
ENSG00000225744.1
ENSG00000266063.1
ENSG00000226516.7
ENSG00000146556.14
ENSG00000217950.4
ENSG00000232742.3
ENSG00000271153.1
ENSG00000204556.4
ENSG00000283443.1
ENSG00000279493.1
ENSG00000277117.5
ENSG00000279687.1
ENSG00000279094.3
ENSG00000274333.4
ENSG00000277777.1
ENSG00000279186.1
ENSG00000274559.3
ENSG00000280013.1
ENSG00000280179.1
ENSG00000274790.1
ENSG00000280019.1
ENSG00000278903.3

ENSG00000274276.4
ENSG00000275895.7
ENSG00000276076.5
ENSG00000280346.1
ENSG00000278927.1
ENSG00000279788.1
ENSG00000280164.1
ENSG00000280018.4
ENSG00000279477.1
ENSG00000277067.4
ENSG00000274046.1
ENSG00000280330.1
ENSG00000279647.1
ENSG00000276077.4
ENSG00000274484.1
ENSG00000277991.4
ENSG00000278961.2
ENSG00000279967.1
ENSG00000275950.1
ENSG00000278233.1
ENSG00000274060.1
ENSG00000277671.1
ENSG00000277739.1
ENSG00000277379.1
ENSG00000274868.1
ENSG00000275215.1
ENSG00000278775.1
ENSG00000279990.1
ENSG00000264462.1
ENSG00000264063.1
ENSG00000224309.7
ENSG00000207097.1
ENSG00000168122.4
ENSG00000185390.2
ENSG00000175302.5
ENSG00000266211.1
ENSG00000226930.1
ENSG00000228184.1
ENSG00000179381.8
ENSG00000166351.11
ENSG00000227874.1
ENSG00000234538.1
ENSG00000228314.1

ENSG00000178457.3
ENSG00000205670.12
ENSG00000222018.2
ENSG00000180509.13
ENSG00000221398.1
ENSG00000142178.9
ENSG00000185186.10
ENSG00000214326.2
ENSG00000160218.13
ENSG00000241945.8
ENSG00000160221.18
ENSG00000237604.1
ENSG00000276871.1
ENSG00000276138.1
ENSG00000184624.4
ENSG00000236831.1
ENSG00000130538.6
ENSG00000235759.1
ENSG00000215268.3
ENSG00000213727.3
ENSG00000230643.1
ENSG00000230471.1
ENSG00000231565.1
ENSG00000226474.1
ENSG00000224435.2
ENSG00000198062.15
ENSG00000225255.6
ENSG00000235992.1
ENSG00000232775.6
ENSG00000275319.1
ENSG00000277690.2
ENSG00000278558.5
ENSG00000274602.5
ENSG00000274625.1
ENSG00000275362.1
ENSG00000286175.1
ENSG00000273907.1
ENSG00000182824.7
ENSG00000197421.10
ENSG00000234764.2
ENSG00000169668.11
ENSG00000239511.2
ENSG00000224688.1

ENSG00000133475.17
ENSG00000283145.1
ENSG00000169662.8
ENSG00000206142.9
ENSG00000226534.1
ENSG00000273846.1
ENSG00000274600.1
ENSG00000200057.1
ENSG00000206140.12
ENSG00000183506.17
ENSG00000252143.1
ENSG00000222352.1
ENSG00000183246.8
ENSG00000206090.4
ENSG00000231271.1
ENSG00000228039.3
ENSG00000099984.11
ENSG00000100280.17
ENSG00000241278.1
ENSG00000251178.1
ENSG00000253540.5
ENSG00000251669.6
ENSG00000227551.1
ENSG00000232399.4
ENSG00000249482.1
ENSG00000250231.1
ENSG00000250844.3
ENSG00000248920.3
ENSG00000249811.3
ENSG00000232264.5
ENSG00000230430.5
ENSG00000227140.3
ENSG00000250566.1
ENSG00000251101.1
ENSG00000184139.8
ENSG00000197465.14
ENSG00000202215.1
ENSG00000261914.2
ENSG00000248308.1
ENSG00000168967.14
ENSG00000250138.4
ENSG00000251158.1
ENSG00000172058.16

ENSG00000172062.17
ENSG00000285204.1
ENSG00000249981.1
ENSG00000250801.2
ENSG00000248943.1
ENSG00000170089.15
ENSG00000214351.5
ENSG00000249287.1
ENSG00000248761.1
ENSG00000231228.4
ENSG00000168903.9
ENSG00000250765.6
ENSG00000213285.4
ENSG00000217929.4
ENSG00000244731.8
ENSG00000204338.8
ENSG00000250535.1
ENSG00000229776.1
ENSG00000272541.1
ENSG00000214563.2
ENSG00000214561.3
ENSG00000211697.4
ENSG00000211698.2
ENSG00000239556.4
ENSG00000228903.7
ENSG00000282879.1
ENSG00000233437.1
ENSG00000225244.3
ENSG00000189166.6
ENSG00000214668.4
ENSG00000275061.2
ENSG00000185177.14
ENSG00000226587.1
ENSG00000229301.1
ENSG00000227305.2
ENSG00000223889.1
ENSG00000268181.3
ENSG00000234467.1
ENSG00000227426.1
ENSG00000224368.1
ENSG00000197990.6
ENSG00000230132.1
ENSG00000228645.2

ENSG00000265214.1
ENSG00000273024.6
ENSG00000230189.7
ENSG00000236928.3
ENSG00000233383.1
ENSG00000174353.17
ENSG00000229018.5
ENSG00000201282.1
ENSG00000273897.1
ENSG00000273927.1
ENSG00000123965.13
ENSG00000239069.1
ENSG00000276840.1
ENSG00000267828.1
ENSG00000273598.1
ENSG00000202021.1
ENSG00000278416.1
ENSG00000199870.1
ENSG00000205583.13
ENSG00000275930.1
ENSG00000233980.1
ENSG00000250614.1
ENSG00000231183.4
ENSG00000241350.1
ENSG00000201885.1
ENSG00000233448.2
ENSG00000201959.1
ENSG00000078319.10
ENSG00000201913.1
ENSG00000170667.15
ENSG00000105808.18
ENSG00000205236.6
ENSG00000267368.1
ENSG00000222011.9
ENSG00000213385.3
ENSG00000237632.3
ENSG00000229977.2
ENSG00000271079.1
ENSG00000252037.1
ENSG00000170379.21
ENSG00000241136.1
ENSG00000198420.10
ENSG00000225932.3

ENSG00000244198.7
ENSG00000212807.3
ENSG00000228960.6
ENSG00000050327.15
ENSG00000241449.6
ENSG00000276538.1
ENSG00000233363.1
ENSG00000253896.3
ENSG00000232039.2
ENSG00000239839.7
ENSG00000284620.2
ENSG00000230045.4
ENSG00000233132.4
ENSG00000249005.3
ENSG00000223885.4
ENSG00000215373.4
ENSG00000233295.3
ENSG00000230549.3
ENSG00000236125.3
ENSG00000215371.3
ENSG00000177257.3
ENSG00000177243.3
ENSG00000187082.2
ENSG00000198129.3
ENSG00000255251.2
ENSG00000248944.1
ENSG00000285975.1
ENSG00000234749.5
ENSG00000285687.1
ENSG00000285814.1
ENSG00000285913.1
ENSG00000285620.1
ENSG00000285937.1
ENSG00000285720.1
ENSG00000285657.1
ENSG00000285607.1
ENSG00000285950.1
ENSG00000285862.1
ENSG00000254776.1
ENSG00000186562.8
ENSG00000176782.3
ENSG00000178287.18
ENSG00000254543.1

ENSG00000255211.1
ENSG00000237038.5
ENSG00000225327.3
ENSG00000233115.4
ENSG00000254229.2
ENSG00000254700.1
ENSG00000255544.2
ENSG00000226430.7
ENSG00000223443.3
ENSG00000205879.5
ENSG00000254423.1
ENSG00000255556.2
ENSG00000251402.3
ENSG00000270971.2
ENSG00000273567.1
ENSG00000205176.3
ENSG00000275928.2
ENSG00000271381.2
ENSG00000276998.1
ENSG00000274343.1
ENSG00000277436.1
ENSG00000226625.1
ENSG00000271454.1
ENSG00000278825.1
ENSG00000269692.1
ENSG00000273994.1
ENSG00000234720.3
ENSG00000232239.2
ENSG00000283886.3
ENSG00000277774.1
ENSG00000232798.1
ENSG00000184961.5
ENSG00000283246.1
ENSG00000275160.1
ENSG00000237207.2
ENSG00000270909.1
ENSG00000273717.1
ENSG00000268234.1
ENSG00000277213.1
ENSG00000234861.2
ENSG00000227924.2
ENSG00000237357.2
ENSG00000204816.5

ENSG00000224537.1
ENSG00000213816.3
ENSG00000276128.2
ENSG00000238933.1
ENSG00000273940.1
ENSG00000229697.2
ENSG00000231995.2
ENSG00000275230.1
ENSG00000227449.8
ENSG00000224603.1
ENSG00000204801.7
ENSG00000236825.2
ENSG00000241571.2
ENSG00000219693.3
ENSG00000170161.7
ENSG00000288694.1
ENSG00000235832.2
ENSG00000277869.1
ENSG00000236252.2
ENSG00000276723.1
ENSG00000275026.1
ENSG00000276203.5
ENSG00000284658.1
ENSG00000276894.1
ENSG00000273849.1
ENSG00000276457.1
ENSG00000274349.5
ENSG00000277490.1
ENSG00000274852.1
ENSG00000276898.1
ENSG00000274098.1
ENSG00000277618.1
ENSG00000232086.1
ENSG00000189357.9
ENSG00000278486.1
ENSG00000265848.1
ENSG00000159247.13
ENSG00000278130.1
ENSG00000268009.6
ENSG00000231603.1
ENSG00000231489.1
ENSG00000274274.1
ENSG00000237671.3

ENSG00000227488.2
ENSG00000236362.8
ENSG00000216649.4
ENSG00000224902.5
ENSG00000189064.8
ENSG00000225397.1
ENSG00000229826.2
ENSG00000269502.6
ENSG00000283594.1
ENSG00000275520.2
ENSG00000269405.7
ENSG00000273746.1
ENSG00000284800.1
ENSG00000215029.11
ENSG00000269226.7
ENSG00000278646.1
ENSG00000226929.3
ENSG00000270493.1
ENSG00000224089.3
ENSG00000271154.1
ENSG00000226600.3
ENSG00000271041.1
ENSG00000230347.5
ENSG00000270397.1
ENSG00000228517.3
ENSG00000271502.1
ENSG00000226023.6
ENSG00000271211.1
ENSG00000237957.6
ENSG00000270736.1
ENSG00000230594.3
ENSG00000270771.1
ENSG00000236126.3
ENSG00000270646.1
ENSG00000242362.2
ENSG00000270456.1
ENSG00000236371.6
ENSG00000270251.1
ENSG00000269096.7
ENSG00000278289.5
ENSG00000271449.8
ENSG00000273696.4
ENSG00000278085.5

ENSG00000269586.8
ENSG00000198573.7
ENSG00000269475.2
ENSG00000268902.3
ENSG00000183305.14
ENSG00000197172.10
ENSG00000277008.1
ENSG00000166160.9
ENSG00000277009.1
ENSG00000269433.3
ENSG00000273041.1
ENSG00000268651.4
ENSG00000184033.14
ENSG00000213338.3
ENSG00000288709.1
ENSG00000221190.1
ENSG00000277745.1
ENSG00000234583.1
ENSG00000235094.3
ENSG00000129816.5
ENSG00000212856.6
ENSG00000237563.2
ENSG00000185700.10
ENSG00000235895.1
ENSG00000232620.1
ENSG00000147761.8
ENSG00000229208.1
ENSG00000252173.1
ENSG00000271595.1
ENSG00000224989.2
ENSG00000283076.1
ENSG00000228578.1
ENSG00000230663.1
ENSG00000252315.1
ENSG00000252012.1
ENSG00000273966.1
ENSG00000238191.3
ENSG00000226449.1
ENSG00000224166.1
ENSG00000229129.1
ENSG00000250868.4
ENSG00000223555.1
ENSG00000235521.1

ENSG00000252513.1
ENSG00000224964.1
ENSG00000225740.1
ENSG00000237616.1
ENSG00000226611.2
ENSG00000236599.1
ENSG00000229302.1
ENSG00000251510.1
ENSG00000238067.1
ENSG00000230377.1
ENSG00000169953.12
ENSG00000236429.1
ENSG00000176728.10
ENSG00000249501.1
ENSG00000248573.1
ENSG00000238135.1
ENSG00000224151.1
ENSG00000251970.1
ENSG00000238088.1
ENSG00000242393.1
ENSG00000242875.6
ENSG00000234414.7
ENSG00000228257.1
ENSG00000242389.8
ENSG00000227444.1
ENSG00000237968.1
ENSG00000231159.1
ENSG00000232475.1
ENSG00000236951.5
ENSG00000131548.7
ENSG00000169800.14
ENSG00000229159.1
ENSG00000228571.1
ENSG00000230476.1
ENSG00000224657.9
ENSG00000224917.1
ENSG00000230727.1
ENSG00000228240.1
ENSG00000276829.1
ENSG00000215537.3
ENSG00000226906.1
ENSG00000244000.1
ENSG00000232764.1

ENSG00000233126.3
ENSG00000215540.10
ENSG00000224169.1
ENSG00000226873.1
ENSG00000236620.1
ENSG00000224866.1
ENSG00000232424.1
ENSG00000252681.1
ENSG00000232910.1
ENSG00000232845.1
ENSG00000231540.1
ENSG00000223362.1
ENSG00000224571.1
ENSG00000226369.1
ENSG00000280961.1
ENSG00000227871.1
ENSG00000251917.1
ENSG00000228465.1
ENSG00000237467.1
ENSG00000232205.1
ENSG00000223698.3
ENSG00000242854.1
ENSG00000230977.1
ENSG00000244231.1
ENSG00000172342.5
ENSG00000284071.1
ENSG00000239533.6
ENSG00000274234.1
ENSG00000225491.1
ENSG00000227439.1
ENSG00000224210.1
ENSG00000235412.1
ENSG00000183795.8
ENSG00000240566.1
ENSG00000232976.1
ENSG00000226270.3
ENSG00000227837.1
ENSG00000278602.1
ENSG00000185894.8
ENSG00000236379.2
ENSG00000227633.1
ENSG00000228786.5
ENSG00000233944.1

ENSG00000270535.1
ENSG00000225287.1
ENSG00000223856.1
ENSG00000223406.1
ENSG00000235511.1
ENSG00000234888.1
ENSG00000235014.1
ENSG00000252948.1
ENSG00000237801.6_PAR_Y
ENSG00000237040.6_PAR_Y
ENSG00000124333.16_PAR_Y
ENSG00000228410.6_PAR_Y
ENSG00000124334.17_PAR_Y
ENSG00000270726.6_PAR_Y
ENSG00000185203.12_PAR_Y
ENSG00000182484.15_PAR_Y
ENSG00000243485.5
ENSG00000284332.1
ENSG00000233750.3
ENSG00000237094.12
ENSG00000269732.1
ENSG00000284733.2
ENSG00000268663.1
ENSG00000284662.1
ENSG00000229376.3
ENSG00000274764.6
ENSG00000204505.5
ENSG00000275131.3
ENSG00000275585.2
ENSG00000273516.1
ENSG00000264145.2
ENSG00000252105.1
ENSG00000162825.16
ENSG00000280649.2
ENSG00000270629.6
ENSG00000274265.5
ENSG00000274423.1
ENSG00000275229.1
ENSG00000203814.6
ENSG00000276997.4
ENSG00000233771.2
ENSG00000233003.4
ENSG00000278455.1

ENSG00000276805.2
ENSG00000228882.1
ENSG00000233867.2
ENSG00000227679.1
ENSG00000228570.8
ENSG00000274599.2
ENSG00000276964.1
ENSG00000275610.1
ENSG00000278664.1
ENSG00000278761.1
ENSG00000278790.1
ENSG00000276904.1
ENSG00000278641.1
ENSG00000253826.1
ENSG00000270921.1
ENSG00000230851.1
ENSG00000255415.1
ENSG00000256136.1
ENSG00000177586.6
ENSG00000270496.1
ENSG00000257162.6
ENSG00000255769.7
ENSG00000257109.3
ENSG00000259454.1
ENSG00000183793.14
ENSG00000277920.1
ENSG00000214940.8
ENSG00000277647.1
ENSG00000276548.1
ENSG00000185864.17
ENSG00000277041.1
ENSG00000198156.10
ENSG00000255524.8
ENSG00000278078.1
ENSG00000169203.16
ENSG00000205534.6
ENSG00000198064.13
ENSG00000260311.1
ENSG00000282927.1
ENSG00000277349.1
ENSG00000275954.5
ENSG00000273513.1
ENSG00000274419.6

ENSG00000278299.6
ENSG00000278599.6
ENSG00000274611.4
ENSG00000188755.10
ENSG00000267792.1
ENSG00000266971.1
ENSG00000282059.1
ENSG00000234837.1
ENSG00000196862.10
ENSG00000183054.12
ENSG00000231943.9
ENSG00000287165.1
ENSG00000221055.1
ENSG00000240356.6
ENSG00000286267.1
ENSG00000286032.1
ENSG00000286148.1
ENSG00000252199.1
ENSG00000214319.2
ENSG00000226444.2
ENSG00000236097.1
ENSG00000237299.1
ENSG00000233408.1
ENSG00000273643.1
ENSG00000278008.1
ENSG00000244699.1
ENSG00000230304.1
ENSG00000251155.2
ENSG00000281720.1
ENSG00000281058.1
ENSG00000280799.1
ENSG00000280757.1
ENSG00000281627.1
ENSG00000280457.1
ENSG00000179978.11
ENSG00000254701.3
ENSG00000145736.14
ENSG00000251634.2
ENSG00000215630.6
ENSG00000230178.1
ENSG00000251035.1
ENSG00000220212.1
ENSG00000243307.2

ENSG00000136206.4
ENSG00000271466.1
ENSG00000273520.5
ENSG00000262461.6
ENSG00000262648.1
ENSG00000184616.9
ENSG00000286038.1
ENSG00000286137.1
ENSG00000286014.1
ENSG00000185040.13
ENSG00000186645.10
ENSG00000205482.10
ENSG00000253612.1
ENSG00000176269.4
ENSG00000214268.2
ENSG00000255025.1
ENSG00000177306.4
ENSG00000233176.3
ENSG00000254889.1
ENSG00000254715.3
ENSG00000254817.1
ENSG00000232948.1
ENSG00000252535.1
ENSG00000250794.2
ENSG00000236875.3
ENSG00000227518.5
ENSG00000283921.1
ENSG00000106714.17
ENSG00000215112.6
ENSG00000252724.1
ENSG00000183148.7
ENSG00000184906.11
ENSG00000154529.15
ENSG00000185775.10
ENSG00000238529.1
ENSG00000286506.1
ENSG00000276040.4
ENSG00000252133.1
ENSG00000244081.1
ENSG00000182021.10
ENSG00000228522.2
ENSG00000156755.10
ENSG00000287988.1

ENSG00000204776.6
ENSG00000275030.1
ENSG00000277350.2
ENSG00000274355.1
ENSG00000274820.1
ENSG00000278551.1
ENSG00000286255.1
ENSG00000168757.12
ENSG00000234830.8
ENSG00000233803.9
ENSG00000225560.6
ENSG00000258992.7
ENSG00000238074.5
ENSG00000230066.6
ENSG00000270073.1
ENSG00000223407.1
ENSG00000172352.5
ENSG00000229709.1
ENSG00000225609.1
ENSG00000227867.1
ENSG00000234744.1
ENSG00000223972.5
ENSG00000279928.2
ENSG00000278757.1
ENSG00000223181.1
ENSG00000207205.1
ENSG00000207501.1
ENSG00000206585.1
ENSG00000212157.1
ENSG00000212230.1
ENSG00000222314.1
ENSG00000264204.2
ENSG00000279058.2
ENSG00000204172.12
ENSG00000222225.1
ENSG00000212612.1
ENSG00000258488.2
ENSG00000261401.1
ENSG00000200132.1
ENSG00000222329.1
ENSG00000259916.1
ENSG00000207476.1
ENSG00000283050.2

ENSG00000198237.8
ENSG00000253366.3
ENSG00000253900.1
ENSG00000235558.3
ENSG00000222533.1
ENSG00000284728.1
ENSG00000284698.1
ENSG00000284603.1
ENSG00000284663.1
ENSG00000240240.9
ENSG00000223839.8
ENSG00000277737.3
ENSG00000176115.9
ENSG00000181997.8
ENSG00000188120.16
ENSG00000198312.4
ENSG00000147996.17

Figure 1

Click here to access/download;Figure;figure1.png

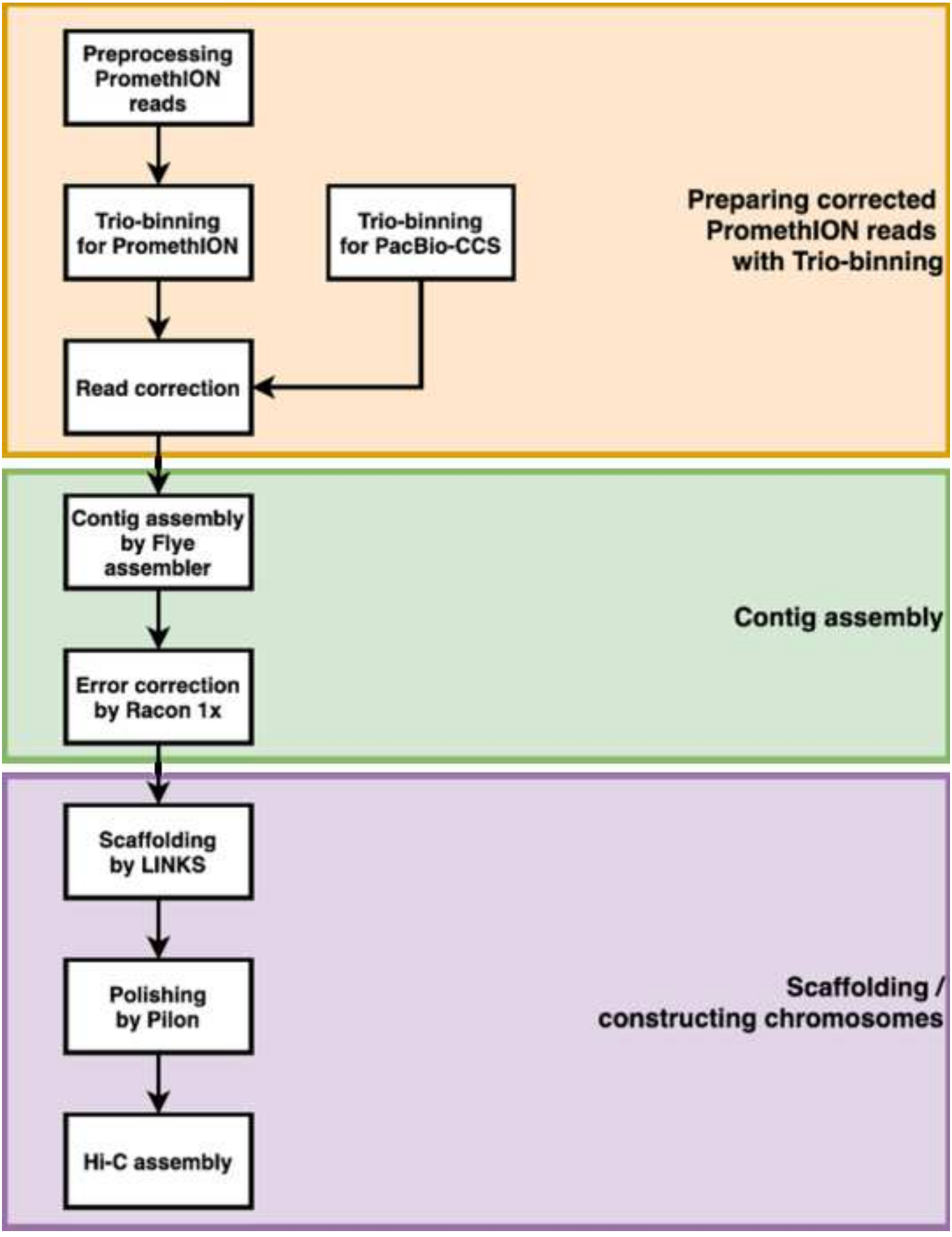

Figure S1

Click here to access/download;Figure;figure_S1.png

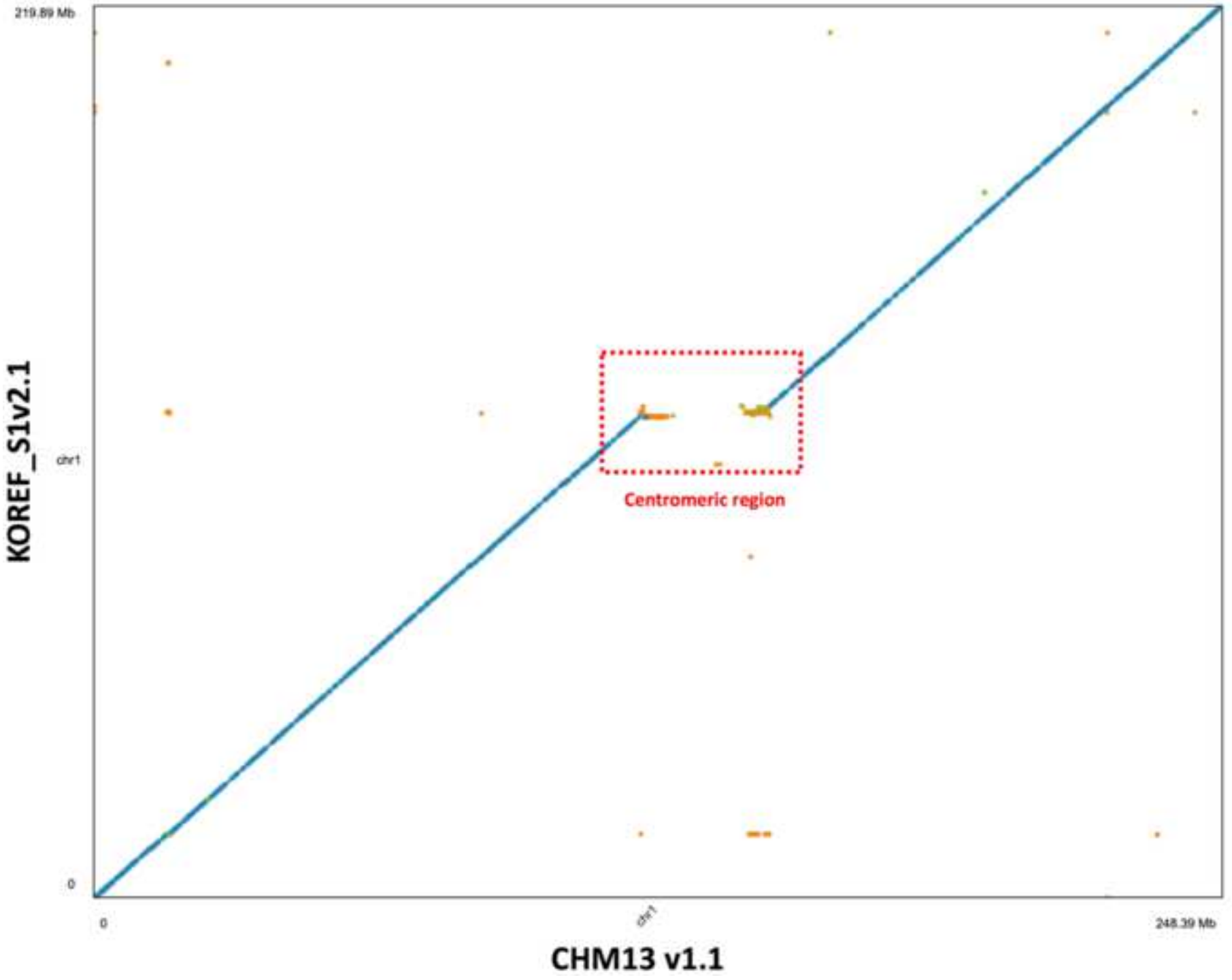

Figure S2

Click here to access/download;Figure;figure_S2.png

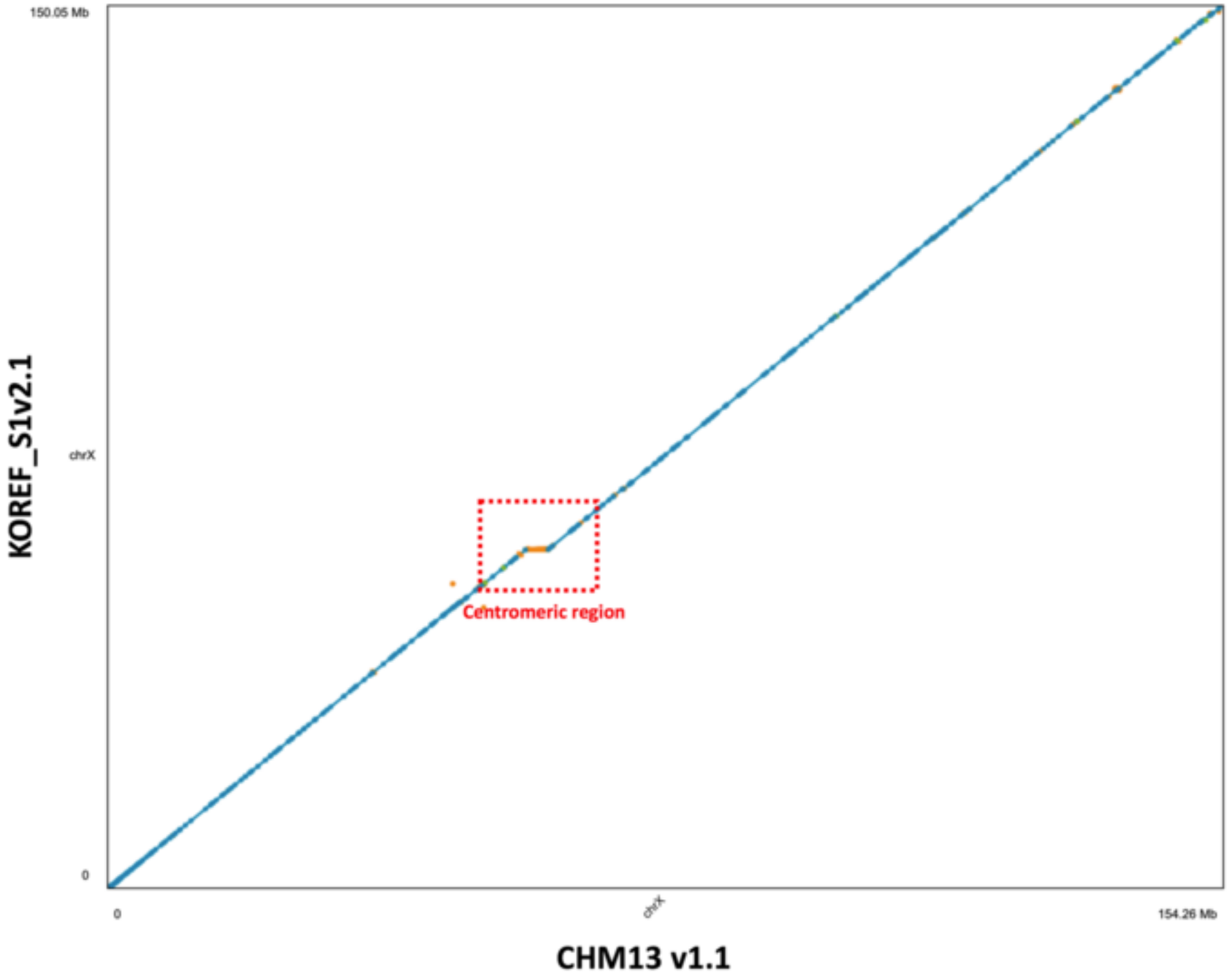

Figure S3a

Click here to access/download;Figure;figure_S3a.png

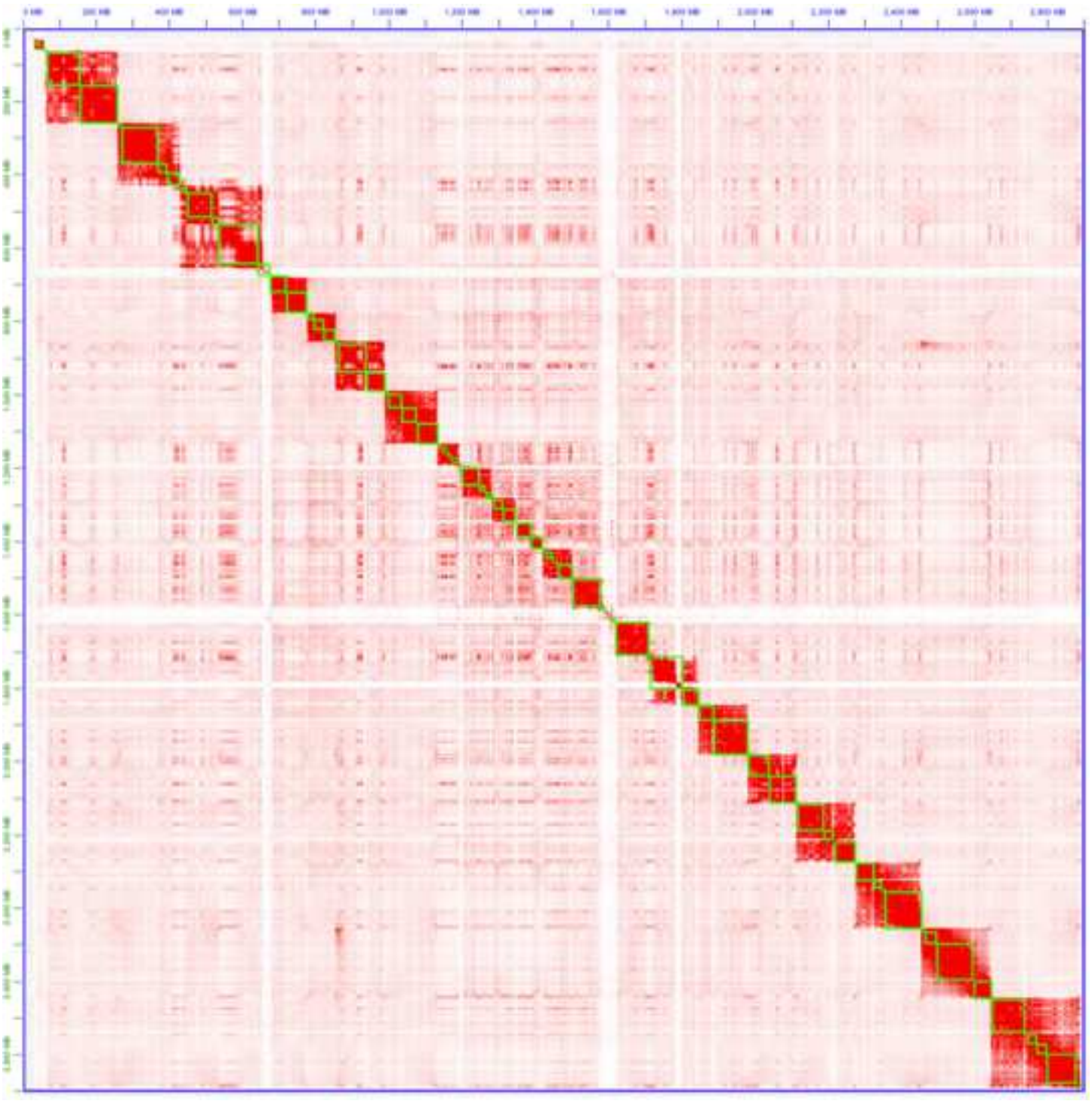

Figure S3b

Click here to access/download;Figure;figure_S3b.png