LETTER

De novo assembly and phasing of a Korean human genome

Jeong-Sun Seo[1,2,3,4,5]*, Arang Rhie[1,2,3]*, Junsoo Kim[1,4]*, Sangjin Lee[1,5]*, Min-Hwan Sohn[1,2,3], Chang-Uk Kim[1,2,3], Alex Hastie[6], Han Cao[6], Ji-Young Yun[1,5], Jihye Kim[1,5], Junho Kuk[1,5], Gun Hwa Park[1,5], Juhyeok Kim[1,5], Hanna Ryu[4], Jongbum Kim[4], Mira Roh[4], Jeonghun Baek[4], Michael W. Hunkapiller[7], Jonas Korlach[7], Jong-Yeon Shin[1,5] & Changhoon Kim[4]

Advances in genome assembly and phasing provide an opportunity to investigate the diploid architecture of the human genome and reveal the full range of structural variation across population groups. Here we report the *de novo* assembly and haplotype phasing of the Korean individual AK1 (ref. 1) using single-molecule real-time sequencing[2], next-generation mapping[3], microfluidics-based linked reads[4], and bacterial artificial chromosome (BAC) sequencing approaches. Single-molecule sequencing coupled with next-generation mapping generated a highly contiguous assembly, with a contig N50 size of 17.9 Mb and a scaffold N50 size of 44.8 Mb, resolving 8 chromosomal arms into single scaffolds. The *de novo* assembly, along with local assemblies and spanning long reads, closes 105 and extends into 72 out of 190 euchromatic gaps in the reference genome, adding 1.03 Mb of previously intractable sequence. High concordance between the assembly and paired-end sequences from 62,758 BAC clones provides strong support for the robustness of the assembly. We identify 18,210 structural variants by direct comparison of the assembly with the human reference, identifying thousands of breakpoints that, to our knowledge, have not been reported before. Many of the insertions are reflected in the transcriptome and are shared across the Asian population. We performed haplotype phasing of the assembly with short reads, long reads and linked reads from whole-genome sequencing and with short reads from 31,719 BAC clones, thereby achieving phased blocks with an N50 size of 11.6 Mb. Haplotigs assembled from single-molecule real-time reads assigned to haplotypes on phased blocks covered 89% of genes. The haplotigs accurately characterized the hypervariable major histocompatability complex region as well as demonstrating allele configuration in clinically relevant genes such as *CYP2D6*. This work presents the most contiguous diploid human genome assembly so far, with extensive investigation of unreported and Asian-specific structural variants, and high-quality haplotyping of clinically relevant alleles for precision medicine.

Although massively parallel sequencing approaches have been widely used to study genomic variation, simple alignment of short reads to a reference genome cannot be used to investigate the full range of structural variation and phased diploid architecture, which are important for precision medicine. By contrast, the single-molecule real-time (SMRT) sequencing platform produces long reads that can resolve repetitive structures effectively. We integrated this technology with several other sequencing approaches to construct a high-quality Korean diploid genome assembly (Extended Data Fig. 1).

SMRT sequencing of the genome of a Korean individual AK1, for whom we have previously reported the annotated variations assessed with BAC clones and array comparative genomic hybridization[1], was performed at 101× coverage using Pacific Biosciences (PacBio) RSII

(Extended Data Fig. 2a). Reads were assembled and error-corrected with FALCON and Quiver[5] to generate 3,128 contigs with a contig N50 length of 17.9 Mb (Extended Data Table 1, Extended Data Fig. 2b and Supplementary Tables 1–3). To anchor these contigs into larger scaffolds, we used next-generation mapping (NGM) from BioNano Genomics Irys System, which produces physical maps with unique sequence motifs that can provide long-range structural information of the genome. Two rounds of NGM at 97× and 108× coverage were performed, with the second designed to protect fragments better from breakage at fragile sites, providing improved long-range anchoring (Supplementary Table 4). The optical maps were assembled *de novo* into genome maps. Hybrid scaffolding of the contigs and genome maps resulted in 2,832 scaffolds with a scaffold N50 size of 44.8 Mb (Extended Data Table 1 and Extended Data Fig. 3a). Because NGMs provide orders of magnitude longer range information (Supplementary Table 4) compared to long reads from the SMRT platform (Supplementary Table 1), we relied on the genome map when there were conflicts between the two datasets. Checks for consistency between genome maps and contigs corrected potential assembly errors within 23 contigs (Extended Data Fig. 3b and Supplementary Table 5). The final assembly after polishing with Illumina reads (Extended Data Fig. 4a) is characterized by marked contiguity that has not been achieved by non-reference assemblies of the human diploid genome[6-8] so far, and improves on the previous best[6] N50 length by 18 Mb (Table 1). The largest 91 scaffolds, for example, cover 90% of the genome and 8 chromosomal arms are spanned by single scaffolds (Fig. 1a).

The scaffolding accuracy of the AK1 assembly was assessed using paired-end sequences from AK1 BAC library[1] from 62,758 BAC clones (Extended Data Fig. 1). Most (95.4%) of the uniquely aligned BAC clones were in concordance with the assembly (Extended Data Table 2), as expected since the genomic DNA originated from the same individual. From the set of BAC clones that aligned concordant with the reference genome, 99.8% also aligned concordant with the AK1 assembly, with most of the discrepancies caused by phase differences (Supplementary Table 6). The base accuracy of the assembly was assessed by Illumina short reads (72×). The read-depth distributions of the reads mapped to GRCh37, GRCh38 and AK1 show similar patterns (Extended Data Fig. 4b). The estimated base-level error rate of the assembly was less than 10^{-5} based on the count of single nucleotide polymorphisms (SNPs) with unexpected alleles (Extended Data Fig. 4c and Supplementary Table 7).

We used the AK1 assembly to close gaps remaining in human genome reference GRCh38. Of 190 euchromatic gaps (Supplementary Table 8), 65 were closed entirely by our *de novo* assembly (Fig. 1b and Extended Data Fig. 5). Local realignment and reassembly, and use of spanning reads, resolved a further 40 gaps. The closed gaps were
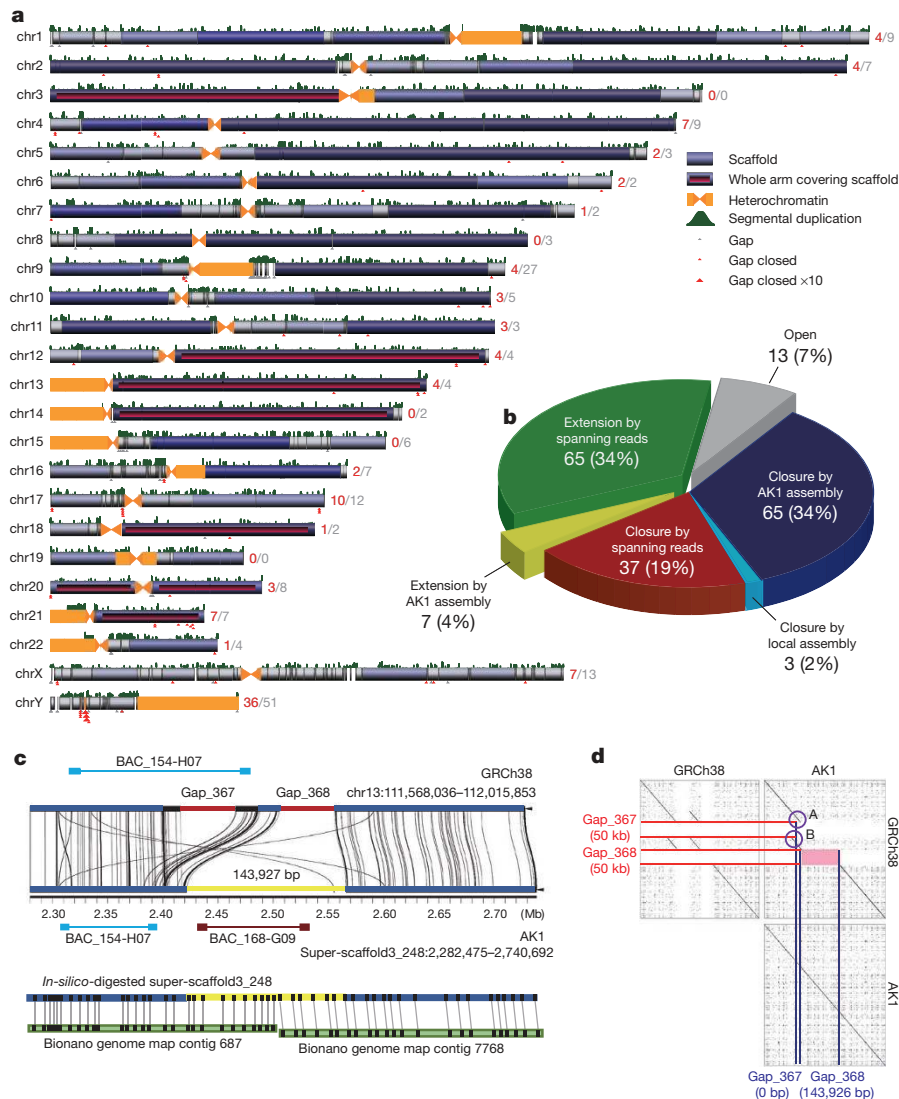
Figure 1 | AK1 *de novo* assembly scaffolds compared to GRCh38.
a, Scaffold coverage over GRCh38 per chromosome. The blue shading represents scaffold size, with darker segments for longer scaffolds. Eight chromosomal arms are spanned by single scaffolds. Closed euchromatic gaps are labelled in red on each chromosome, with the total number of gaps in grey. **b**, Number of gaps closed using the AK1 assembly (blue), local assembly of long reads (light blue), and long reads alone (red). The number of extended gaps with AK1 assembly is represented in yellow, with long reads in green and open gaps in grey. The 65 dot plots of gaps

closed with the AK1 assembly can be found in the AK1 genome browser (http://211.110.34.36/gbrowse2). **c**, AK1 assembly resolving two gaps along with BACs and optical map suggests that gap_367 and both its edges (red and black bars) shrink to zero, whereas gap_368 expands to 144 kb (yellow bar). **d**, Three dot plots show how unique sequences have been added to the reference genome. Reference–reference (top left), reference–AK1 assembly (top right) and AK1–AK1 (bottom right). A and B indicate deleted GRCh38 sequence around gap_367.

filled with a total of 364 kb of sequence into 1.5 Mb (Supplementary Table 9 and Supplementary Information). We also extended into 72 of the 85 remaining gaps with the addition of 663 kb of sequence into 4.1 Mb. These locations, previously intractable using only short reads, commonly contained simple tandem repeats, as reported previously[6,9]. One example (Fig. 1c, d) illustrates two gaps resolved by AK1 assembly with supporting evidence from BACs and genome maps.

We identified 18,210 structural variants (SVs), including 7,358 deletions, 10,077 insertions, 71 inversions, and 704 complex variants at a base resolution through the direct comparison between the AK1 assembly and the human reference genome GRCh37 (Supplementary Tables 10, 11). We were able to validate 271 out of 276 SVs with BAC contigs generated by SMRT sequencing (Supplementary Table 12). Compared to previous studies[6,8–11], a total of 11,927 variants were previously unreported, which account for approximately 47% (3,465) and 76% (7,710) of all deletions and insertions, respectively (Fig. 2a and Extended Data Fig. 6a). Of the new SVs, 86% were highly enriched for

clusters of mobile and tandem repeats (Extended Data Fig. 6b). PacBio long-read sequencing of the corresponding transcriptome revealed that 155 isoforms are expressed from 54 novel insertion loci, indicating the existence of functional elements in human genomes that were probably undetectable using short reads (Supplementary Table 13). A total of 4,326 deletions and 5,833 insertions occurred within 6,073 genes. Out of 615 exonic variants, 427 were new, and 68% of them did not affect protein functionality by maintaining the reading frame or occurring within non-protein coding genes. Among the new amino-acid-changing variants, 77% were composed of mobile or tandem repeats (Supplementary Table 14), and functional annotation clustering with the 31 genes, which contain the remaining non-repetitive variants, using DAVID[12] showed that they were predominantly related to ion binding, epidermal growth factor, and fibronectin.

Investigation of the insertions suggested that the AK1 sequences consist not only of repeats and duplications, but also of unique sequences that are not found in the reference genome. To examine
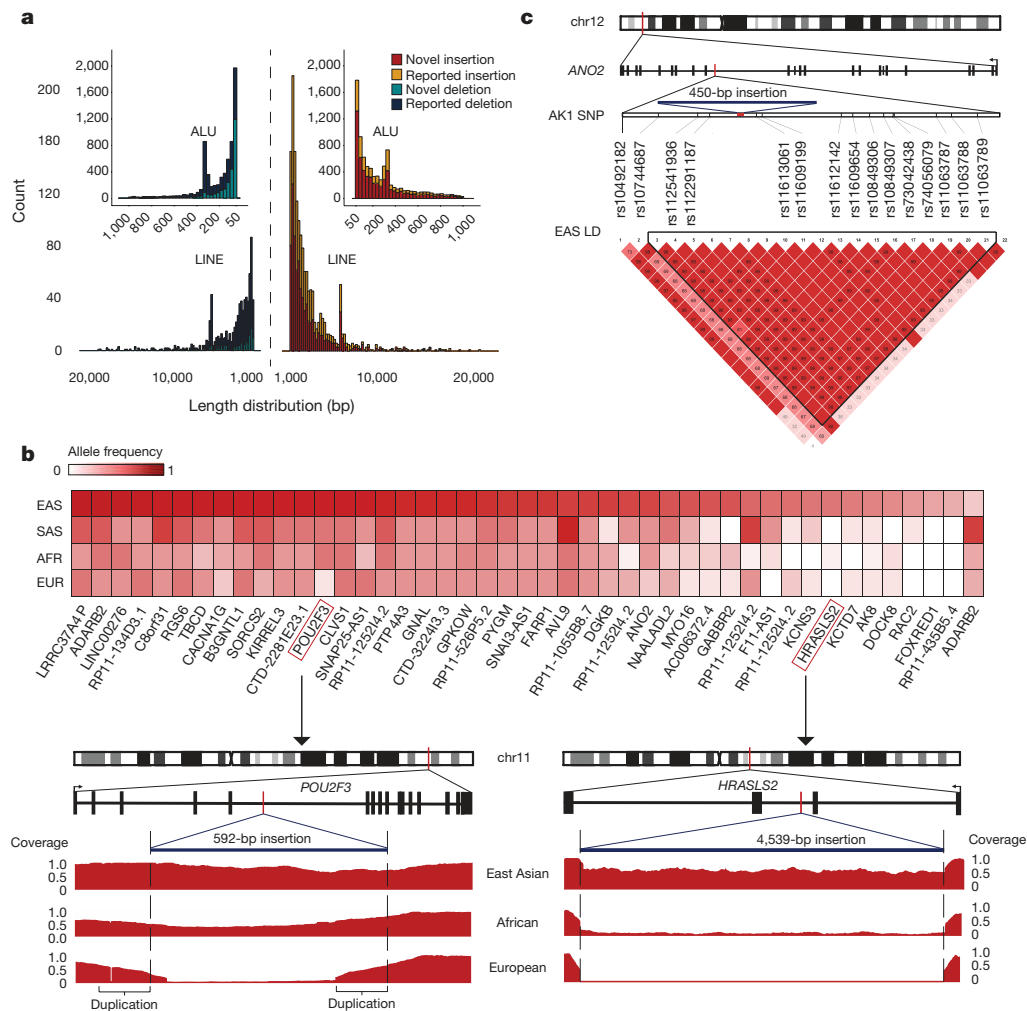
Figure 2 | AK1 SV distribution and Asian-specific variants.
a, Distribution of insertions (red/orange) and deletions (cyan/dark blue) between AK1 and GRCh37, compared to SVs identified from previous studies. In total. 47% and 76% of the insertions and deletions, respectively, were previously unreported. **b**, Allele frequency of 45 Asian-specific insertions ($\geq$0.3 allele frequency difference; $\leq$0.5 non-Asian allele frequency). The coverage for the genic insertions was calculated from 38 whole-genome high-coverage samples by dividing the read depth by the median genome coverage across individuals with the same ancestry. **c**, In *ANO2*, the Asian-specific insertion occurs within an East Asian (EAS) linkage disequilibrium (LD) block, sharing a similar population allele frequency with the adjacent AK1 SNPs. AFR, African; EUR, European; SAS, South Asian.

whether the unique sequences are universal or ancestry specific, we aligned raw reads from high-coverage 1000 Genomes Project samples[10,11] and additional high-coverage Asian samples against our AK1 assembly, and compared the normalized read depths between four ancestral groups. Out of 853 insertions, encompassing 1.7 Mb, which were found in all of the ancestral groups, 800 insertions were also called from the variant analysis with respect to GRCh38, and as such are candidates for addition to the human reference genome (Supplementary Tables 15, 16). Moreover, 400 insertions showed highly polymorphic frequency variability across the populations, and 76 of them, including 45 genic insertions, were Asian specific. Among the genic insertions, we found that a 592-bp insertion within *POU2F3*, reported to have distinctly variable haplotype frequencies among populations[13], was comprised of 452 bp of unique sequence between two 140-bp duplications (Fig. 2b). We also identified numerous large insertions with higher frequency in the Asian population, such as a 4,539-bp insertion in *HRASLS2*. Next, we investigated the haplotype structures associated with Asian-specific variants by using linkage disequilibrium blocks inferred from 1000 Genomes Project Asian samples[10]. Among the variants, 39 insertions were present within the blocks, and 82% of them were located on the same block as a homozygous AK1 SNP, the frequency of which was highest in the Asian

population (Supplementary Table 17). One of the insertions, found within *ANO2*, had a similar allele frequency with adjacent homozygous AK1 SNPs within the same linkage disequilibrium block, suggesting that the insertion shares a single ancient haplotype with the SNPs (Fig. 2c). Our findings demonstrate the important genomic differences of Asian ancestral group from the others, and highlight the need for further genomic studies focused on individuals outside of European ancestry to describe the full range of functionally important variations in humans.

To reflect the diploid genome structure better, we built separate *de novo* assemblies (haplotigs) representing the two haplotypes of each homologous chromosome pair[14]. Phasing was performed with PacBio long reads, Illumina short reads, 10X Genomics linked reads[4] (30$\times$), and reads from BACs representing a single haplotype (47$\times$). Heterozygous SNVs called from these methods are unambiguously assigned to two alternative phases, producing phased blocks with an N50 length of 11.6 Mb, considerably longer than previously reported[4,6,8,15,16] (Table 1). We assessed the accuracy of the phased blocks against the end sequences of BACs, and found a long-range switch error rate to be under 0.3%. SMRT reads were then partitioned into the two phases in which sufficient marker SNVs were present. The two partitioned read sets were assembled *de novo* into haplotigs (Table 1 and Extended Data Table 3).

Table 1 | Comparison of human *de novo* assembly and haplotype phasing summary statistics

	AK1	HuRef	YH_2.0	NA12878	GRCh38
Assembly approach	WGS and BAC	WGS	WGS and fosmid	WGS	BAC and fosmid
Sequencing and physical mapping	PacBio and BioNano	Sanger	Illumina and CG	PacBio and BioNano	Sanger, FISH, OM and fingerprint contigs
De novo assembly algorithm	FALCON	Celera	SOAPdenovo2	Celera and FALCON	Multiple methods
Phasing approach	*De novo*	Reference-guided	*De novo*	Reference-guided	NA
Scaffold/contig N50 (Mb)	44.85/17.92	17.66/0.11	20.52/0.02	26.83/1.56	67.79/56.41
Scaffold/contig L50	21/50	48/7,164	39/40,005	37/532	16/19
No. of scaffolds/contigs	2,832/4,206	4,530/71,333	125,643/361,157	18,903/21,235	735/1,385
No. of gaps	264*	68,109†	235,514†	2,332*	999†
Total gap length (Mb)	37.34	34.43	105.20	146.35	159.97
Total bases/non-N bases in assembly (bp)	2,904,207,288 /2,866,687,809	2,844,000,504 /2,809,571,127	2,911,235,363 /2,806,031,133	3,176,574,379 /3,030,222,093	3,209,286,105 /3,049,316,098
Phased block N50 (Mb)	11.55	0.35	NA	0.15	NA
No. of haplotigs	18,964	NA	24,597	NA	NA
Haplotig N50 (kb)	875	NA	484	NA	NA
Haplotig sum (bp)	4,804,460,182	NA	5,152,727,603	NA	NA

We compared the sequencing platform, algorithms, assembly and phasing statistics of human assemblies so far. The comparison demonstrates the power of single-molecule technologies to generate assemblies with superior assembly statistics than that achieved by short-read sequencing. The assembly statistics were obtained from the NCBI and if the summary statistics were not available from NCBI, the numbers were directly acquired from relevant papers. The accession numbers for HuRef[7], YH_2.0 (ref. 8), NA12878 (ref. 6) and GRCh38 assemblies are GCA_000002125.2, GCA_000004845.2, GCA_001013985.1 and GCA_000001405.15, respectively. CG, complete genomics; FISH, fluorescent *in-situ* hybridization; NA, not applicable; OM, optical mapping; WGS, whole-genome shotgun.
*Number of spanned gaps.
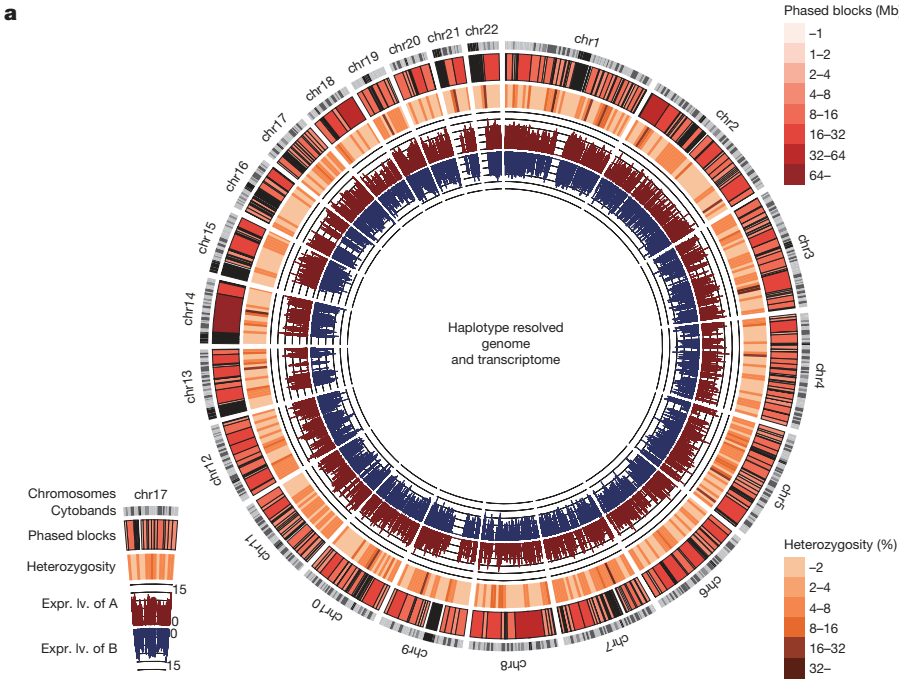†Number of spanned and unspanned gaps.

Figure 3 | Circular visualization of phased blocks with phase-specific expression and two phased regions of MHC class II and *CYP2D6*. **a**, Genome-wide map of highly heterozygous regions and expression levels of haplotype A and B in log scale. **b**, HLA genes in the MHC class II region. This highly variable, complex region contained many SVs, making it difficult to phase against the reference genome, but allowed full resolution through the *de novo* approach. For detailed comparison, see Extended Data Fig. 7. **c**, Both haplotypes of *CYP2D6* and *CYP2D7*. A duplicated copy of *CYP2D6* was fused with the last exon of *CYP2D7* on haplotype B.
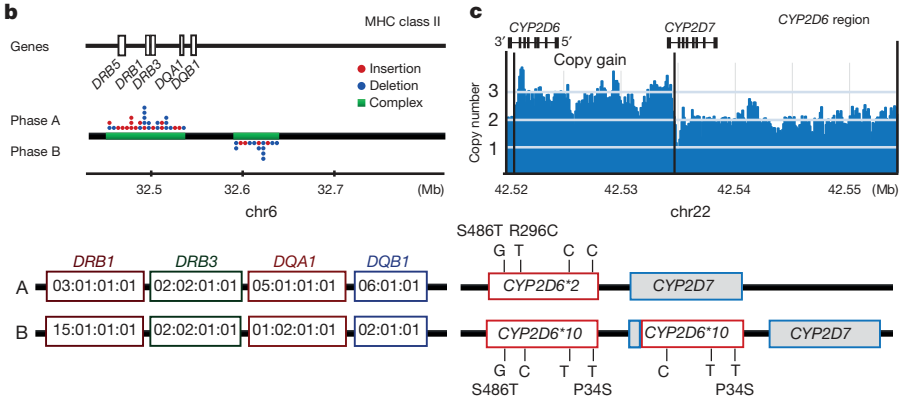
Comparison of the haplotigs to the human reference led to identification of haplotype-specific alleles including SNPs, short indels and SVs (Supplementary Table 18). In addition to the SVs called from the assembly, 13,436 heterozygous haplotype-specific SVs were identified from haplotigs. We tested the accuracy of these SVs against BAC contigs on the same phase, and found that 67 out of the 69 that could be assessed matched perfectly (Supplementary Table 19). The combined length of SNVs, indels and SVs that were heterozygous between the two haplotigs was 69.8 Mb. Moreover, we were able to measure the expression level from each haplotype genome widely (Fig. 3a).

We examined the haplotypes of human leukocyte antigen (HLA) genes in detail, and confirmed the haplotypes using targeted SMRT sequencing (Supplementary Table 20). To avoid common problems[17] associated with hyperpolymorphic patterns of allelic variation, major histocompatibility complex (MHC) class I and II regions were assembled independently. The MHC class II region was phased successfully despite a large number of SVs, highlighting the utility of our *de novo* phasing approach (Fig. 3b and Extended Data Fig. 7). Our approach also allowed a clinically important duplication of *CYP2D6* to be detected and assigned to one phase (Fig. 3c). This result demonstrates that *de novo* assembly-based phasing has advantages in resolving challenging hypervariable regions, and could be used further for pharmacogenomics (Supplementary Discussion).

Allelic configuration is also particularly important for recessive traits. For example, we were able to phase two genes that contained more than two nonsynonymous, heterozygous alleles known to be associated with recessive diseases (Supplementary Table 21). Variants in *MEFV*[18] and *ADAMTS13* (ref. 19), which are predicted to cause familial Mediterranean fever and Upshaw–Shalman syndrome under the autosomal recessive inheritance pattern, respectively, were found in *cis* configuration, with the partner haplotype left intact.

These results demonstrate the power of *de novo* genome assembly and phasing by integrating SMRT sequencing, genome maps, linked reads and BACs for the generation of high-quality contiguous scaffolds, the detection of the full range of SVs, and for understanding the haplotype structure in clinically relevant genes for precision medicine.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Kim, J.-I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460,** 1011–1015 (2009).
2. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
3. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30,** 771–776 (2012).
4. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34,** 303–311 (2016).
5. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).
6. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12,** 780–786 (2015).
7. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
8. Cao, H. *et al.* De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33,** 617–622 (2015).
9. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).
10. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
11. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
12. Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4,** 44–57 (2009).
13. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343,** 1017–1021 (2014).
14. Makoff, A. J. & Flomen, R. H. Detailed analysis of 15q11-q14 sequence corrects errors and gaps in the public access sequence to fully reveal large segmental duplications at breakpoints for Prader–Willi, Angelman, and inv dup(15) syndromes. *Genome Biol.* **8,** R114 (2007).
15. Suk, E.-K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21,** 1672–1685 (2011).
16. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).
17. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16,** 627–640 (2015).
18. Bernot, A. *et al.* Non-founder mutations in the *MEFV* gene establish this gene as the cause of familial Mediterranean fever (FMF). *Hum. Mol. Genet.* **7,** 1317–1325 (1998).
19. Kokame, K. *et al.* Mutations and common polymorphisms in *ADAMTS13* gene responsible for von Willebrand factor-cleaving protease activity. *Proc. Natl Acad. Sci. USA* **99,** 11902–11907 (2002).

Supplementary Information is available in the online version of the paper.

Author Contributions J.-S.S. and C.K. conceived and designed the experiments. J.-Y.S., J.-Y.Y. and Ji.Ki. conducted sequencing and relevant experiments. J.Ku., G.H.P. and Juh.Ki. performed BAC clone library preparation and sequencing. A.H. and H.C. generated BioNano data. H.R. performed RNA-seq and isoform sequencing analysis. Jo.Ki. performed phasing of 10X Genomics linked reads. M.R. and J.B. performed *de novo* assembly. J.Ko. and M.W.H. performed PacBio sequencing. M.-H.S. performed gap closure. Jun.Ki., S.L. and C.-U.K. performed SV analysis. A.R. performed phasing analysis. J.-S.S., A.R., Jun.Ki., S.L., M.-H.S. and C.K. primarily wrote the manuscript, although many authors provided edits.

Author Information The accession codes for the underlying sequence data are summarized in Supplementary Table 22. The AK1 assembly and haplotigs have been deposited at DDBJ/ENA/GenBank under the accessions LPVO00000000 and LYWJ00000000, respectively. The AK1 assembly, haplotigs, BAC placements and SVs from various platforms and dot plots of gaps resolved by the AK1 assembly are available as a browsable track on AK1 genome browser (http://211.110.34.36/gbrowse2). Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.-S.S. (jeongsun@snu.ac.kr) or C.K. (kimchan@macrogen.com).

Reviewer Information *Nature* thanks S. Salzberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

AK1 cell line. An immortalized lymphoblastoid cell line was established from the AK1 individual through Epstein–Barr virus transformation of mononuclear cells (Seoul Clinical Laboratories Inc.). Full pathogen testing was performed and maintained in a mycoplasma-free facility. AK1 lymphoblastoid cell line was cultured in RPMI 1640 media containing 15% FBS at 37 °C in a humidified 5% CO_2 environment. The approval number C-0806-023-246 for the AK1 individual was assigned based on the guidelines from the Institutional Review Board of Seoul National University.

PacBio data generation. Genomic DNA was extracted from AK1 cells using the Gentra Puregene Cell Kit (Qiagen). Large-insert PacBio library preparation was conducted following the Pacific Biosciences recommended protocols. In brief, a total of 60 μg AK1 genomic DNA was sheared to ~20 kb targeted size by using Covaris g-TUBEs (Covaris). Each shearing processed 10 μg input DNA and a total of 6 shearings were performed. The sheared genomic DNA was examined by Agilent 2100 Bioanalyzer DNA12000 Chip (Agilent Technologies Inc.) for size distribution and underwent DNA damage repair/end repair, blunt-end adaptor ligation followed by exonuclease digestion. The purified digestion products were loaded onto pre-cast 0.6% agarose for 7–50 kb size selection using the BluePippin Size Selection System (Sage Science), and the recovered size-selected library products were purified using 0.5× pre-washed Agencourt AMPure XP beads (Beckman Coulter). The final libraries were examined by Agilent 2100 Bioanalyzer DNA12000 Chip for size distribution and the library concentration was determined with Qubit 2.0 Fluorometer (Life Technologies). We sequenced with the PacBio RSII instrument with P6 polymerase binding and C4 chemistry kits (P6C4). A total of 380 SMRT Cells were used to yield 101-fold whole-genome sequence data.

Sample preparation for BioNano Genomics. AK1 cells were pelleted and washed with PBS; the final cell pellet was re-suspended in cell-suspension buffer using the CHEF Mammalian Genomic DNA Plug Kit (Bio-Rad). Cells were then embedded in CleanCut low-melt Agarose (Bio-Rad) and spread into a thin layer on a custom support (in development at BioNano Genomics). Cells were lysed using IrysPrep Lysis Buffer (BioNano Genomics), protease-treated with Puregene Proteinase K (Qiagen), followed by brief washing in Tris with 50 mM EDTA and then washing in Tris with 1 mM EDTA before RNase treatment with Puregene RNase (Qiagen). DNA was then equilibrated in Tris with 50 mM EDTA and incubated overnight at 4 °C before extensive washing in Tris with 0.1 mM EDTA followed by equilibration in NEBuffer 3 (New England BioLabs) at 1× concentration. Purified DNA in the thin layer agarose was labelled following the IrysPrep Reagent Kit protocol with adaptations for labelling in agarose. In brief, 1.25 μg of DNA was digested with 0.7 U Nt.BspQI nicking endonuclease per microlitre of reaction volume in NEBuffer 3 (New England BioLabs) for 130 min at 37 °C, then washed with TE Low EDTA Buffer (Affymetrix), pH 8.0, followed by equilibration with 1× ThermoPol Reaction Buffer (New England BioLabs). Nick-digested DNA was then incubated for 70 min at 50 °C using the IrysPrep Labelling mix (BioNano Genomics) and Taq DNA Polymerase (New England BioLabs) at a final concentration of 0.4 U μl^{-1}. Nick-labelled DNA was incubated for 40 min at 37 °C using the IrysPrep Repair mix (BioNano Genomics) and Taq DNA Ligase (New England BioLabs) at a final concentration of 1 U μl^{-1}. Labelled-repaired DNA was then recovered from the thin layer agarose by digesting with GELase and counterstained with IrysPrep DNA Stain (BioNano Genomics) before data collection on the Irys System. The fragile site rescue process protects fragile sites by reducing the temperature of the labelling reaction and minimizes shear forces by restraining DNA in agarose until nicks are repaired. In this case, only the closest opposite-strand nick-pairs break.

Sequencing library preparation using the GemCode platform. Sample indexing and partition barcoded libraries were prepared using GemCode Gel Bead and Library Kit (10× Genomics)[4]. Sequencing was conducted with Illumina Hiseq2500 to generate linked reads.

Illumina data generation. Libraries were generated with PCR-free protocols. gDNA was sheared twice using Covaris S2 with cycling conditions of 10% duty cycle, Cycles/Burst 200, and Time 100 s. The sheared DNA was processed using the Illumina TruSeq DNA PCR-Free LT Library Kit protocol to generate 550 bp inserts, which includes end repair, SPRI bead size selection, A-tailing, and Y-adaptor ligation. Library concentration was measured by qPCR and loaded on HiSeq X Ten instruments (PE-150) to generate 72-fold sequence coverage.

DNA preparation from BAC clones. A total of 32,026 BAC clones were selected from the 252 384-well plates and re-plated into 96-well plates. Clones were grown overnight, and the cultures were used to prepare two additional replicates for the two 384-well plates that were stored at −80 °C in LB medium containing 20%

glycerol. A total of 32,026 clone cultures with growth at ODs ranging from 0.6 to 1.0 were pooled, pelleted and the DNA was extracted using the standard alkaline lysis method. In this procedure, a cell pellet was resuspended in 150 μl of Qiagen buffer P1 with RNase and lysed with 150 μl of 0.2 M NaOH, 1% SDS solution for 5 min. Lysis was neutralized with the addition of 150 μl of 3 M sodium acetate, pH 4.8. Neutralized lysate was incubated on ice for 30 min, and DNA was collected by centrifugation for 15 min at 15.7g at 4 °C, concentrated by standard ethanol precipitation and resuspended in 25 μl of 10 mM Tris-HCl, pH 8.5.

PacBio sequencing of BAC clones. DNA from approximately 150 BAC clones with roughly equimolar concentration was combined into a single pool. A total of 10 μg from each pool DNA was sheared and fragments of insert size ranging from 10 to 15 kb were selected. Two libraries were prepared from the pooled DNA using a PacBio SMRTbell library preparation kit v1.0. The libraries were quantified using a Qubit 2.0 fluorometer and each library was sequenced using two SMRT cells with P6C4 chemistry.

Illumina sequencing of BAC clones. DNA from approximately 290 BAC clones with roughly equimolar concentration was combined into a single BAC pool. One nanogram of DNA from each pool was digested and fragments of insert size ranging from 500 to 550 bp were selected. In total, 109 libraries were prepared from the pooled DNA using Illumina-compatible Nextera XT DNA sample prep kit and sequenced with HiSeq2500.

Sample preparation for RNA sequencing. We extracted RNA from tissue using RNAiso Plus (Takara Bio), followed by purification using RNeasy MinElute (Qiagen). RNA was assessed for quality and was quantified using RNA 6000 Nano LabChip on a 2100 Bioanalyzer (Agilent). The RNA sequencing (RNA-seq) libraries were prepared as previously described[20]. RNA library was sequenced with Illumina TruSeq SBS Kit v3 on a HiSeq 2000 sequencer (Illumina) to obtain 100 bp paired-end reads. The image analysis and base calling were performed using the Illumina pipeline with default settings.

Sample preparation for isoform sequencing. Total RNA extracted from AK1 cells with RNA integrity number (RIN) > 8.0 was used for library preparation. The library was constructed following the Clontech SMARTer-PCR cDNA Synthesis Sample Preparation Guide. 1–2 kb, 2–3 kb, 3–6 kb and >5 kb libraries were selected by Sage, ELF purified, end-repaired and blunt-end SMRTbell adapters were ligated. The fragment size distribution was confirmed on a Bioanalyzer HS chip (Agilent) and quantified on a Qubit fluorometer (Life Technologies). The fragment size distribution was validated on a Bioanalyzer HS chip (Agilent) and quantified on a Qubit fluorometer (Life Technologies). The sequencing was carried out on the PacBio RSII instrument using P6C4.

PacBio long-read *de novo* assembly. Around 31 million subreads were used for assembly with FALCON v0.3.0 (ref. 21) given length_cutoff parameter of 10 kb for initial mapping to build pre-assembled reads (preads), and preads over 15 kb were used (length_cutoff_pr) to maximize the assembled contig N50 (Extended Data Fig. 2). Primary and associated contigs were polished using Quiver[5].

BioNano Genomics genome map generation. Optical maps were *de novo* assembled into genome maps using BioNano assembler software (Irys System, BioNano Genomics). Single molecules longer than 150 kb with at least 8 fluorescent labels were used to find possible overlaps ($P < 1 \times 10^{-10}$). Next, these maps were constructed to consensus maps by recursively refining and extending them by mapping single molecules ($P < 1 \times 10^{-5}$). The consensus maps were compared and merged into genome maps when patterns matched ($P < 1 \times 10^{-10}$). A second set of optical maps was obtained thereafter, and generated into genome maps with the same criteria.

Contig editing and hybrid assembly. Primary contigs were *in silico* digested into cmaps and were compared with genome maps for scaffolding. The scaffolding was visualized and performed with the Irys Viewer. When conflict occurred, the contigs were edited with the guidance of genome map.

Assembly improvements. Paired-end reads from Illumina platform were aligned to the assembly using bwa[22] mem, followed with duplication removal using Picard tools[23]. Base-pair correction of the assembly was performed using Pilon[24]. Pilon mostly corrected single insertions and deletions in regions enriched with homopolymer. Contigs or scaffolds shorter than 10 kb were excluded from the overall analysis to avoid results from spurious misassembly.

Scaffold accuracy measurement with BAC clones. Scaffolding accuracy of the AK1 assembly was assessed using the AK1 BAC library[1]. AK1 BAC end sequences (BES) were aligned to GRCh37, GRCh38 and AK1 assemblies using BWA. The BES placements were categorized by the alignment, orientation and separation of BES with respect to the assembly. The BES placement was determined to be concordant: (1) if the BES placement was placed in the same assembly unit; (2) if the paired end sequences were properly oriented; and (3) if the *in silico* insert size was between 50,000 and 250,000 bp. If the BES placements did not meet these conditions, the BES placement was defined to be discordant. In addition, if only one

of the paired end sequences were aligned to the assembly, the BES placement was defined to be an orphan placement. If both paired-end sequences were unaligned to the assembly, the BES was defined to be unmapped. If either of the paired-end sequences were aligned to different positions of the assembly multiple times, the BES was defined to have multiple placements.

Gap closure and SV analysis: alignment to the reference genome. To identify the precise genomic location of each assembly unit, we used LASTZ[25] with parameters (-gapped -gap = 600,150,-hspthresh = 4500,-seed = 12of19 -notransition -ydrop = 15000-chain) to align each assembly unit to each chromosome in the human reference genome. Chaining procedure was followed to join the neighbouring local alignments into a single cohesive alignment. The chained alignments of each assembly unit were processed to obtain a single alignment with the best alignment score. If the selected alignment was not fully representative of the assembly unit, we selected a set of alignments that was better representative of the assembly unit. A netting procedure was then followed with the selected chained alignments. The chaining and netting procedures were applied using UCSC Kent tools[26] and parallel processing was used when possible to increase computational speed.

Gap closure of GRCh38. Gaps were classified into telomeric, centromeric, heterochromatic, acrocentric and euchromatic region according to the agp file and cytoband information provided by the Genome Reference Consortium (GRC) and UCSC genome browser. In total, 190 euchromatic gaps were targeted for gap closure with AK1 assembly. The gaps that could not be closed or extended with the AK1 assembly were subjected to closure through local assembly using Canu[27] or a contiguous subread. Subreads mapped 10 kb upstream or downstream of the gap were chosen for local assembly. Alignment was performed with BLASR[28] -bestn 3, and primary aligned reads with mapping quality of 254 were used. The assembled contigs were thereafter aligned to their respective gap position to precisely identify the added sequences. Subreads used to close the gaps were chosen following criteria described in the Supplementary Information.

Assembly based variant detection. The alignments of the assembly to the reference genome were parsed to obtain SNPs, indels and SVs, which we defined as insertion, deletion, inversion and complex variants with event size equal to or greater than 50 bp. The complex SVs are the same as 'double-sided insertion' defined previously[29]. We used GRCh37 instead of GRCh38 for the main analysis for compatibility and comparison with previously reported structural variations.

SV annotation. Repeat elements were annotated using RepeatMasker (-species human -no_is) and tandem repeat finder (TRF) (2 7 7 80 10 50 2000 -f -m -h -d). SVs are classified accordingly if it is masked by at least 70% with a single type. Complex is defined as the SVs having either several annotated repeat elements, or at least 30% of the remaining sequence not annotated as repeat. Novelty was identified by comparing the breakpoints with 50% reciprocal overlap criterion. Functional annotation was performed using both GENCODE release v19 (GRCh37) and v21 (GRCh38)[30] and the Ensembl Regulatory Build[31]. For those SVs that occurred within gene regulatory domains, we annotated with the nearest gene name. SV located within pericentromeric regions (5 Mb flanking annotated centromeres) and subtelomeric regions (150 kb from the annotated telomeric sequence) were annotated as heterochromatin. Both pilot and strict accessibility genome mask regions (version 20141020) were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/. Segmental duplication sites were downloaded from the UCSC table browser. To simplify categorization of the SVs that lie within multiple functional regions, they were classified according to the order of priority as follow: coding sequence, untranslated region, intron, transcription-factor-binding site, promoter, enhancer, CTCF (transcriptional repressor), and intergenic. To annotate whether the SVs called from GRCh37 were also shared with GRCh38 SV sets, we compared each AK1 breakpoints with 50% reciprocal overlap criterion. In addition, we assessed whether the SVs called from GRCh38 were also represented in the alternative contigs by measuring the concordance against the SV regions including the surrounding 50 bp from the breakpoints.

Asian-specific SVs. Population allele frequency of SVs was obtained by aligning reads from 38 high-coverage samples from five different ancestral backgrounds (African, American, European, East Asian, and South Asian) to the AK1 assembly. We obtained whole-genome sequencing data of 23 individuals from the 1000 Genomes Project and we additionally sequenced 15 East Asian individuals (5 Japanese, 5 Chinese and 5 Koreans). Analysis candidates were selected from the insertions with less than 70% of repeats. We excluded any duplications among the insertions that are mapped to GRCh37 using BLAST (-evalue 1e-10 -perc_identity 90 -qcov_hsp_perc 90). The regions that have been recognized as mobile element or tandem repeat by RepeatMasker and TRF softwares were masked for analysis. Normalized read depth within the unique sequence was achieved by dividing the read depth, which was calculated using samtools bedcov, by the

median genome coverage. The insertions were determined to be highly polymorphic if there were greater than or equal to 0.3 variant frequency differences across the different populations. Asian-specific insertions were chosen by selecting the insertions with equal or above 0.3 allele frequency difference between Asian and non-Asian population as well as non-Asian allele frequency with equal or below 0.5. Asian linkage disequilibrium blocks were obtained from East Asian samples in the 1,000 Genomes Project phase 3 using S-MIG++ algorithm[32] (-maf 0.05 -ci AV -probability 0.95). Linkage disequilibrium blocks with below 0.8 haplotype diversity index were excluded.

***De novo* phasing markers.** We performed phasing against the *de novo* assembly. SNPs and short indels called from whole-genome sequencing (72×) of short reads were phased with linked reads. The non-redundant set of PacBio subreads were aligned to the assembly, and corrections were applied by calculating the maximum likely variant allele for the phased variants based on the read depth. A phased block was defined as the region spanning two markers which had a subread or linked read information providing phasing. Similar to the linked reads, Illumina sequenced BAC phase information was used to correct phasing markers and extend phased blocks. Correction and other bioinformatics methods were performed using an in-house script, described in the Supplementary Information.

Switch error of phasing markers. Long-range switch error measurements were obtained using BAC end sequences. The end sequences were aligned to the AK1 assembly with bwa mem, and the base allele of the phasing marker site was called with the corresponding BAC information. When switching occurred for more than two marker sites in a phased block, it was defined as a long range switch. The long-range switch error rate was calculated as: no. of long range switches/no. of phasing markers.

Haplotig assembly. Using the final set of phasing markers, subreads were classified into sets of haplotype A or B when >85% of the phasing markers agreed. When a subread contained no marker, it was classified as homozygous. Through the read depth, phasing markers that were missed in previous steps were additionally called for homozygous regions adjacent to phased blocks. Subreads in haplotype A or homozygous regions were assembled into haplotig A, and haplotype B into haplotig B with Canu[27]. Haplotigs for MHC class I and II were assembled separately to avoid misassemblies owing to high sequence homology between HLA genes. In this case, subreads phased as homozygous were used with subreads of haplotype A and B. Homozygously phased subreads flanked on each side of a sequencing gap belonged on haplotype A and B, respectively, and were re-classified during assembly.

Haplotype-specific variant calling and annotation. Haplotype-specific variants were called following the assembly-based variation calling method. Owing to possibilities of false positives introduced by misassembly, phased variants that agreed with initial variants called with whole genome sequencing reads were used for further analysis. After functional annotation using GENCODE v19 (ref. 30), disease risk alleles were screened using ClinVar[33]. Haplotyping of *CYP2D6* was done by comparing haplotigs to M33388 following *CYP2D6* nomenclature.

***De novo* assembly of BAC clones and SV validation.** BACs identified to be discordant in size (>1 kb) were pooled and sequenced with the SMRT platform. The subreads were assembled using Canu[27] after screening and removing *Escherichia coli* or vector sequences with CrossMatch[34]. The assembled BAC contigs were polished with Quiver. The BAC contigs were, thereafter, used to validate AK1 assembly-based or phase-specific SVs by assessing the concordance between the assembly and the BAC contig at sites of detected SVs.

Heterozygosity and allele-specific expression. On the basis of the alignments of haplotigs to GRCh37, haplotig A and B were localized to compare partner sequences. The number of different bases were summed in every 5 Mb distance, and percentiled to draw in the Fig. 3a. RNA-seq reads were trimmed and aligned to GRCh37 using STAR aligner[35] with the two-pass mapping strategy as recommended. Duplicates were removed using Picard tools, and variants were called using HaplotypeCaller and VariantFiltration following GATK best practices on RNA-seq[36]. Sites with supportive evidence of altered variation in RNA-seq have been extracted from the final vcf file, and ASEReadCounter[37] was applied to remove reads with low base quality. Read counts are annotated to the phase-specific variants called from haplotigs using in-house scripts. When read depth for one allele was over 30, it was considered as 'expressed'.

20. Seo, J.-S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.* **22,** 2109–2119 (2012).
21. PacificBiosciences/FALCON; GitHub, available at: https://github.com/PacificBiosciences/FALCON (accessed 2 August 2016)
22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
23. Picard Tools by Broad Institute; available at: http://broadinstitute.github.io/picard (accessed 2 August 2016)

24. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).
25. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA.* PhD thesis, Penn State Univ. (2007).
26. ENCODE-DCC/kentUtils; GitHub, available at: https://github.com/ENCODE-DCC/kentUtils (accessed 2 August 2016)
27. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Preprint at http://dx.doi.org/10.1101/071282 (2016).
28. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13,** 238 (2012).
29. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100,** 11484–11489 (2003).
30. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
31. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).
32. Taliun, D., Gamper, J., Leser, U. & Pattaro, C. Fast sampling-based whole-genome haplotype block recognition. *IEEE/AMC Trans. Comput. Biol. Bioinf.* **13,** 315–325 (2016).
33. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
34. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
35. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
36. The GATK Best Practices for variant calling on RNAseq, in full detail; available at: http://gatkforums.broadinstitute.org/wdl/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail (accessed 2 August 2016).
37. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16,** 195 (2015).
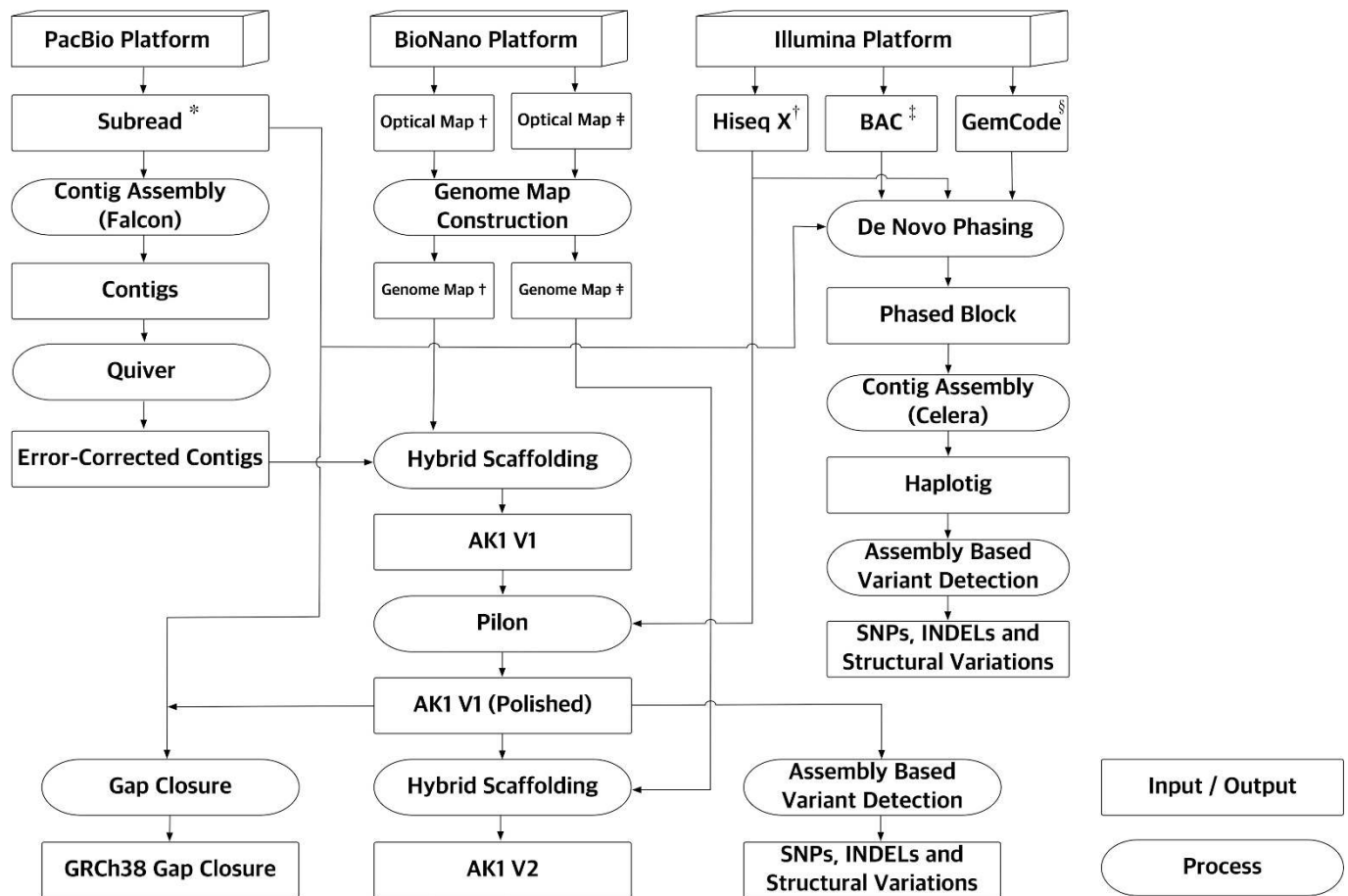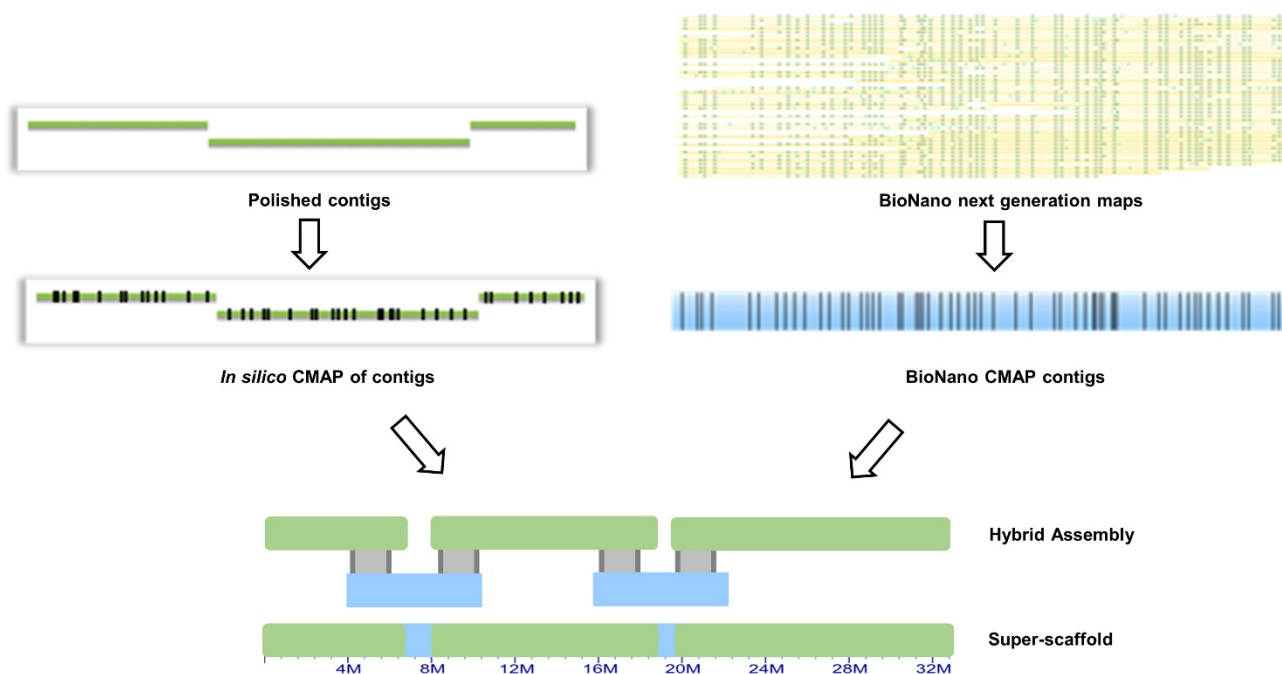
Extended Data Figure 1 | Global overview of data generation and sequencing throughput. Flowchart of the data generation, processing and analysing for the *de novo* assembly and haplotype phasing of the AK1 diploid genome. ∗The SMRT platform sequencing throughput is described in Supplementary Table 1. †The number of read and sequencing throughputs from the Illumina platform are 1,635,192,864 and 249,914,122,464 bp, respectively. ‡AK1 BAC library was sequenced using Sanger capillary end sequencing (single end: 22,563, paired end: 62,758), Illumina (31,719) and SMRT (307) platform. §Linked-read data were additionally generated with the GemCode platform to produce 1,153,598,732 reads from high molecular mass DNA with an average insert size of 100 kb.
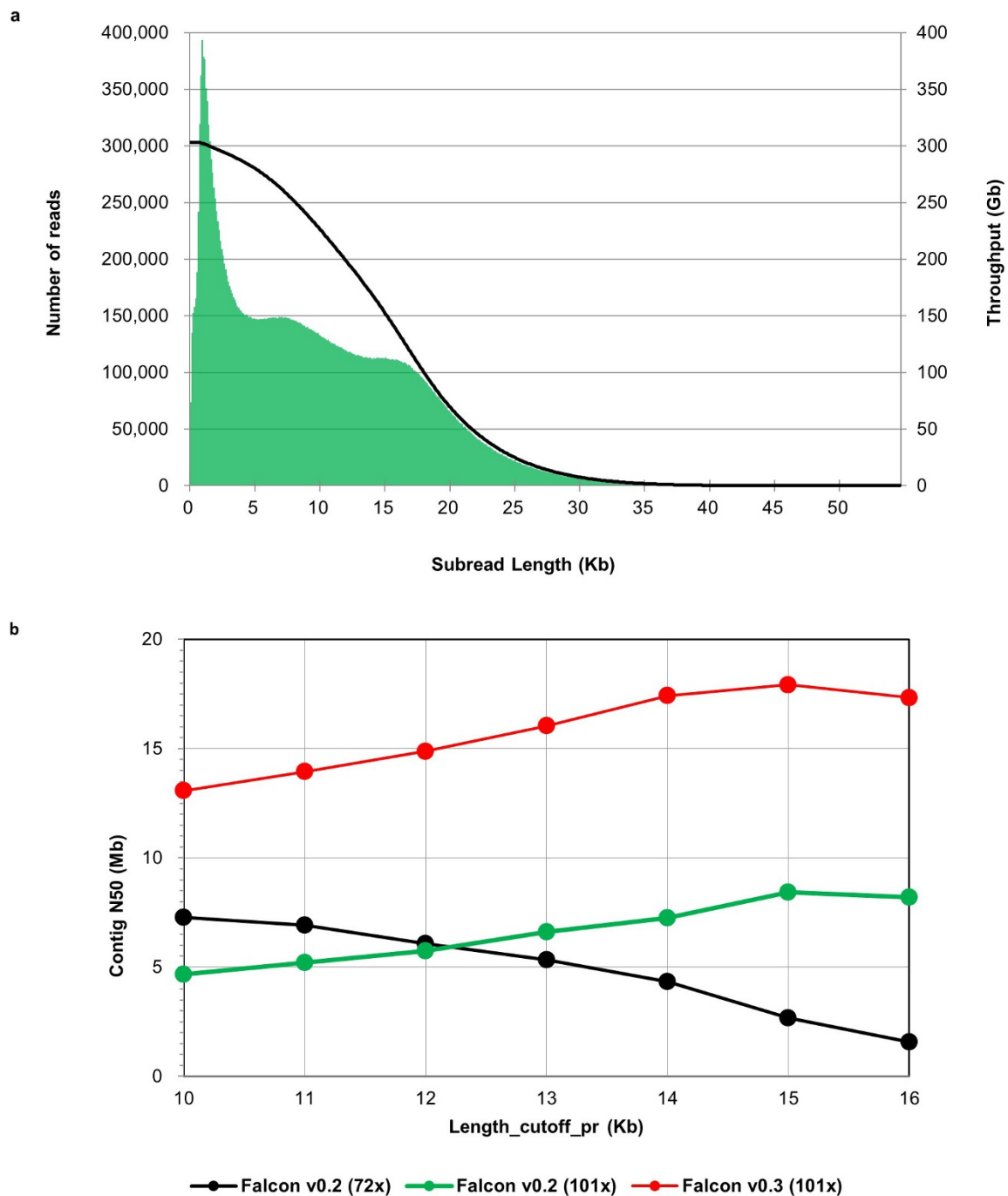
Extended Data Figure 2 | Length distribution of SMRT subreads and FALCON parameter optimization for assembly. a, The *y* axis on the left shows the number of subreads with given length (bin size = 100 bp) on the *x* axis, whereas the *y* axis on the right shows the sum of the length of subreads longer than or equal to the given length on the *x* axis. **b**, Effects of length cutoff parameters on contig N50 in *de novo* assembly by FALCON is shown on the right. The contig N50 depends on the two parameters, related to the amount of error-corrected reads for final assembly, length_cutoff and length_cutoff_pr, respectively, where the former was fixed at 10 kb but the latter varied from 10 to 16 kb. Black and green lines indicate the changes of N50 for 72× and 101× sequencing dataset, respectively.

a

Polished contigs

BioNano next generation maps

In silico CMAP of contigs

BioNano CMAP contigs

Hybrid Assembly

Super-scaffold

4M   8M   12M   16M   20M   24M   28M   32M

b

2-way split on contig 000029F

3-way split on contig 000251F
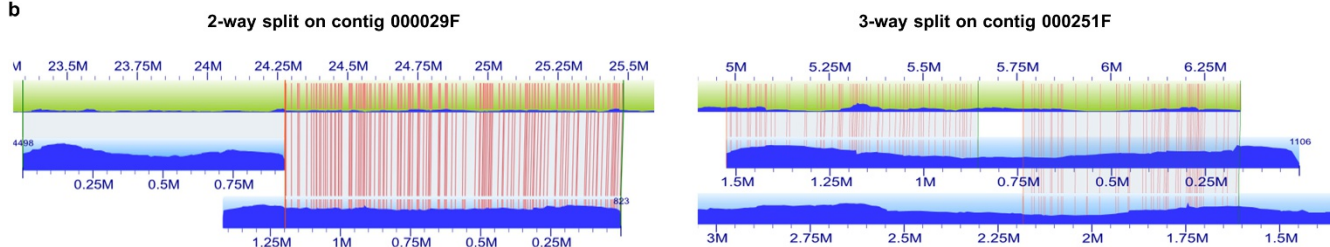
Extended Data Figure 3 | Graphical representation of hybrid assembly and statistics for next generation map and genome map. a, The hybrid assembly approach aligns *in silico* generated maps from sequence contigs with genome maps. When genome maps bridge two contigs, a scaffold is produced. The comp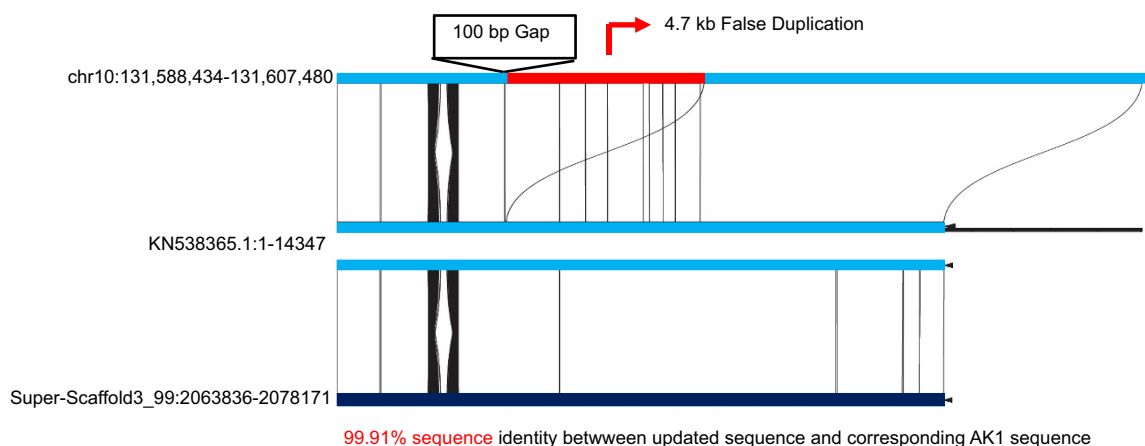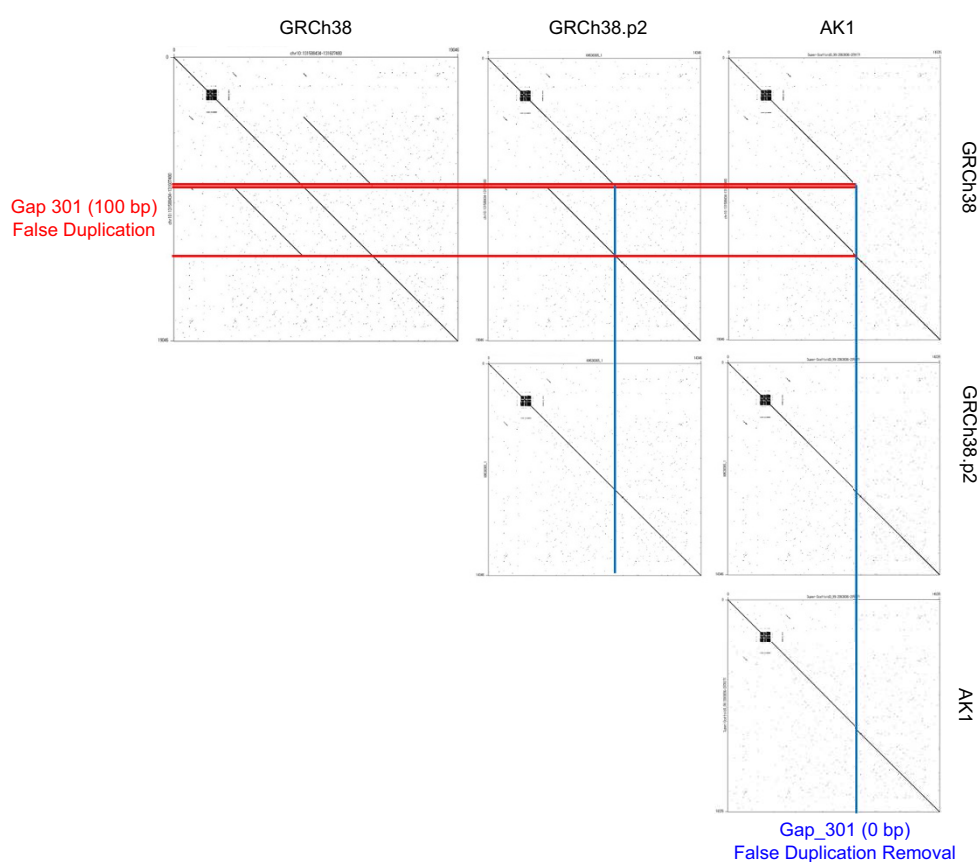arison is visualized between the genome maps and contigs in the Iris Viewer. **b**, Examples of edited contigs due to conflicts between the contig and the genome maps. The matches between the *in silico* map and the genome map are highlighted in red, and mismatches are indicated by absence of the red lines.
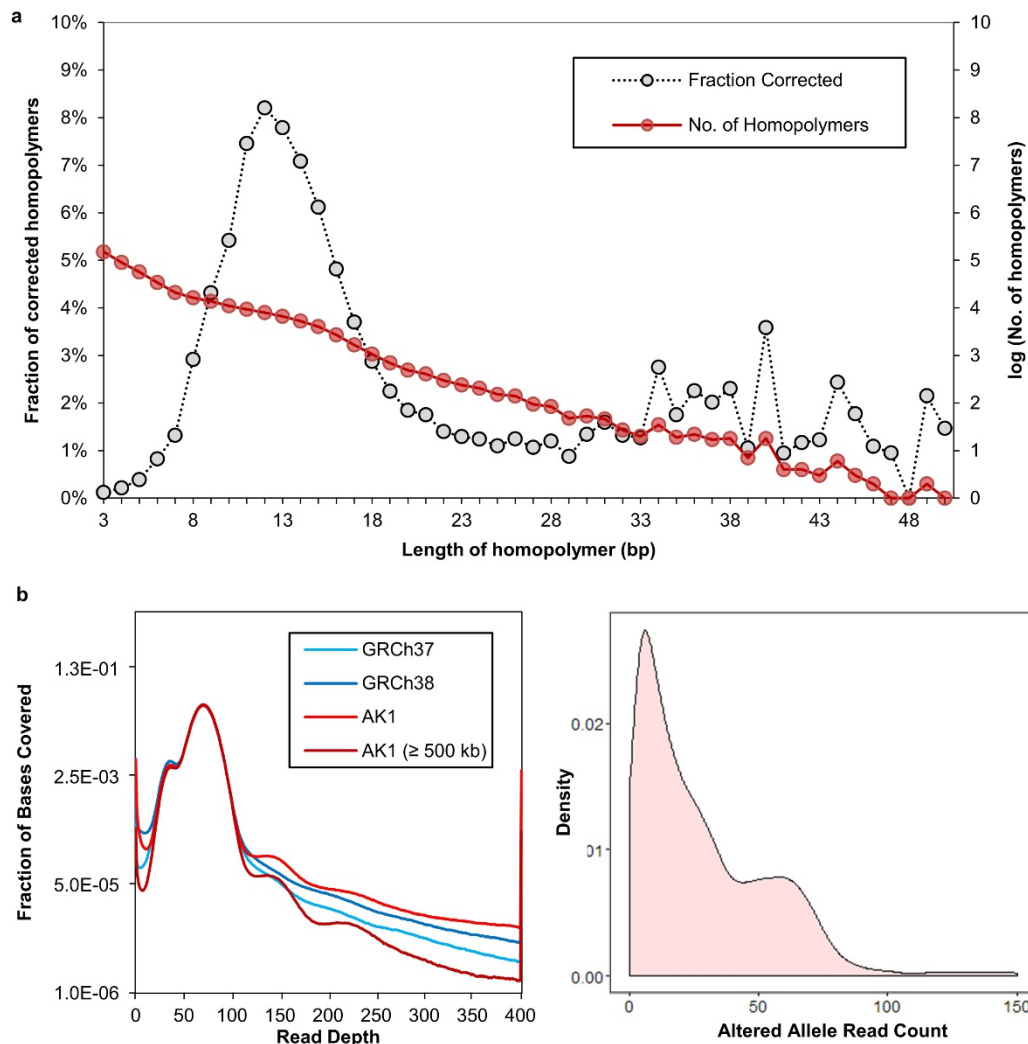
Extended Data Figure 4 | Assessment of assembly accuracy with homopolymer and read depth coverage generated with short reads.
a, Distribution of corrections in homopolymer. Pilon mostly corrected the single base deletions in the assembly and the corrections are enriched in regions with long stretches of homopolymer. **b**, The read-depth distribution against AK1 assembly, AK1 assembly with scaffolds ≥ 500 kb, GRCh37 and GRCh38. As the mean coverage depth of short reads was $72\times$, a peak is shown around it representing the fraction of autosomal region. Another peak is shown in $\sim 36\times$, which is half of the mean coverage depth, representing the contigs derived from chromosomes X and Y. The fluctuating long tale is showing 3-copy and 4-copy of a haplotype, but more clearly observed with AK1 long scaffolds. The overall pattern is showing that more SVs are reflected in AK1 long contigs than the reference. The short contigs (<500 kb) are only 120.4 Mb, comprising a small fraction of the AK1 assembly. **c**, Density plot of the homozygously altered allele read depth from long scaffolds (≥ 500 kb). Most variants are skewed in low allelic read depth, suggested to be mainly due to sequencing artefact or mapping bias.

a

100 bp Gap

4.7 kb False Duplication

chr10:131,588,434-131,607,480

KN538365.1:1-14347

Super-Scaffold3_99:2063836-2078171

99.91% sequence identity betwween updated sequence and corresponding AK1 sequence

b

GRCh38   GRCh38.p2   AK1

Gap 301 (100 bp)
False Duplication

GRCh38

GRCh38.p2

AK1

Gap_301 (0 bp)
False Duplication Removal

Extended Data Figure 5 | An example of filled sequence that matches perfectly with the patch sequence (KN538365.1). a, One AK1 scaffold (Super-scaffold3_99) closes a 100-bp gap in chromosome 10, reducing the size of this gap to zero while it also removes a 4.7 kb false duplication found left of the gap. This information corresponds perfectly to the GRCh38 fix patch (KN538365.1) sequence covering this region, thus validating our assembly and gap closing accuracy. **b,** Six dot plots show the comparison between GRCh38, KN538365.1 and the AK1 assembly. The dot plots are organized in the following manner: Reference-reference (top left), KN538365.1-reference (top middle), AK1-reference (top right), KN538365.1-KN538365.1 (centre middle), AK1-KN538365.1 (middle right) and AK1-AK1 (bottom right).

a

b
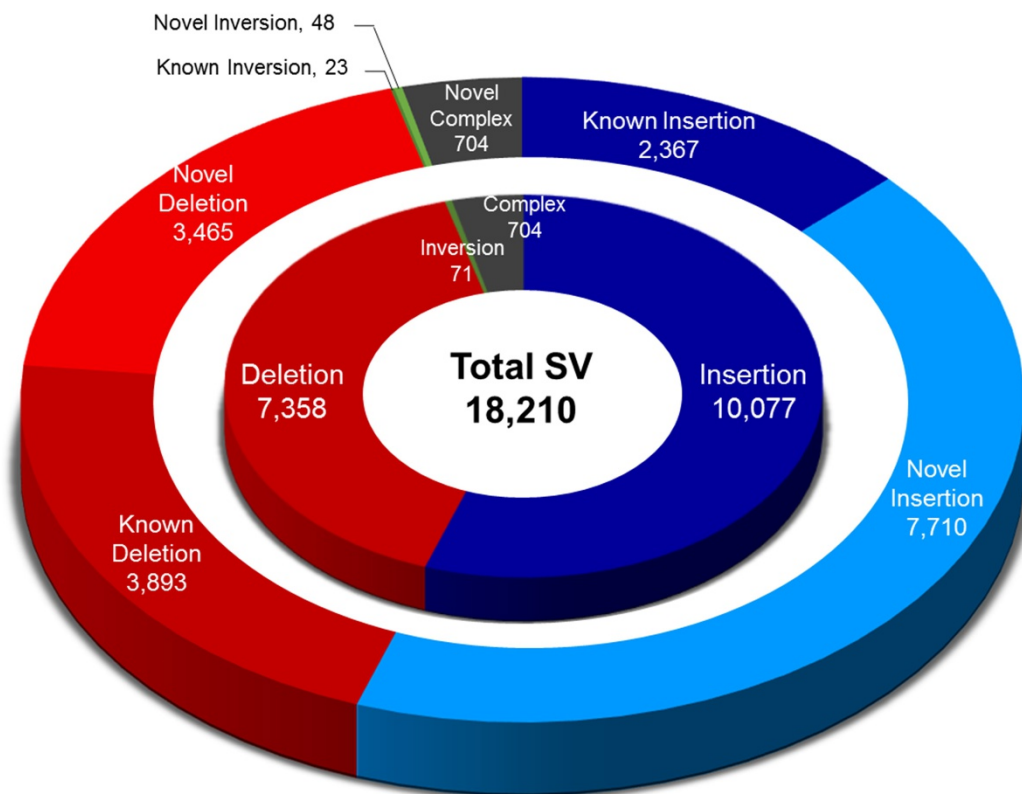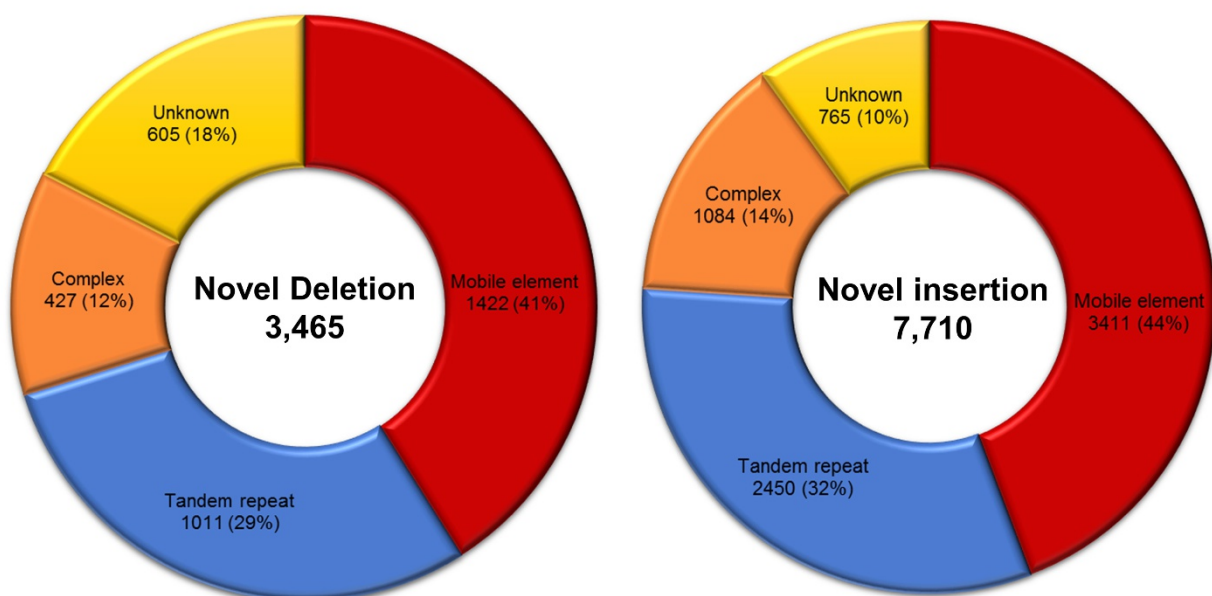
Extended Data Figure 6 | Number of SVs and repeat composition.
a, Overall distribution of SVs. By direct comparison between AK1 assembly and GRCh37 reference genome, deletion (red), insertion (blue), inversion (green), and complex (grey) variants were detected. Outer pie chart represents new variants for each SV types. In total, 65% (11,927)

of the SVs were unreported previously. **b**, Repeat composition of AK1 insertion and deletion. Both insertions and deletions are mostly composed of mobile elements or tandem repeats. Complex is defined as the SVs having either several annotated repeat elements, or at least 30% of the remaining sequence not annotated as repeat.
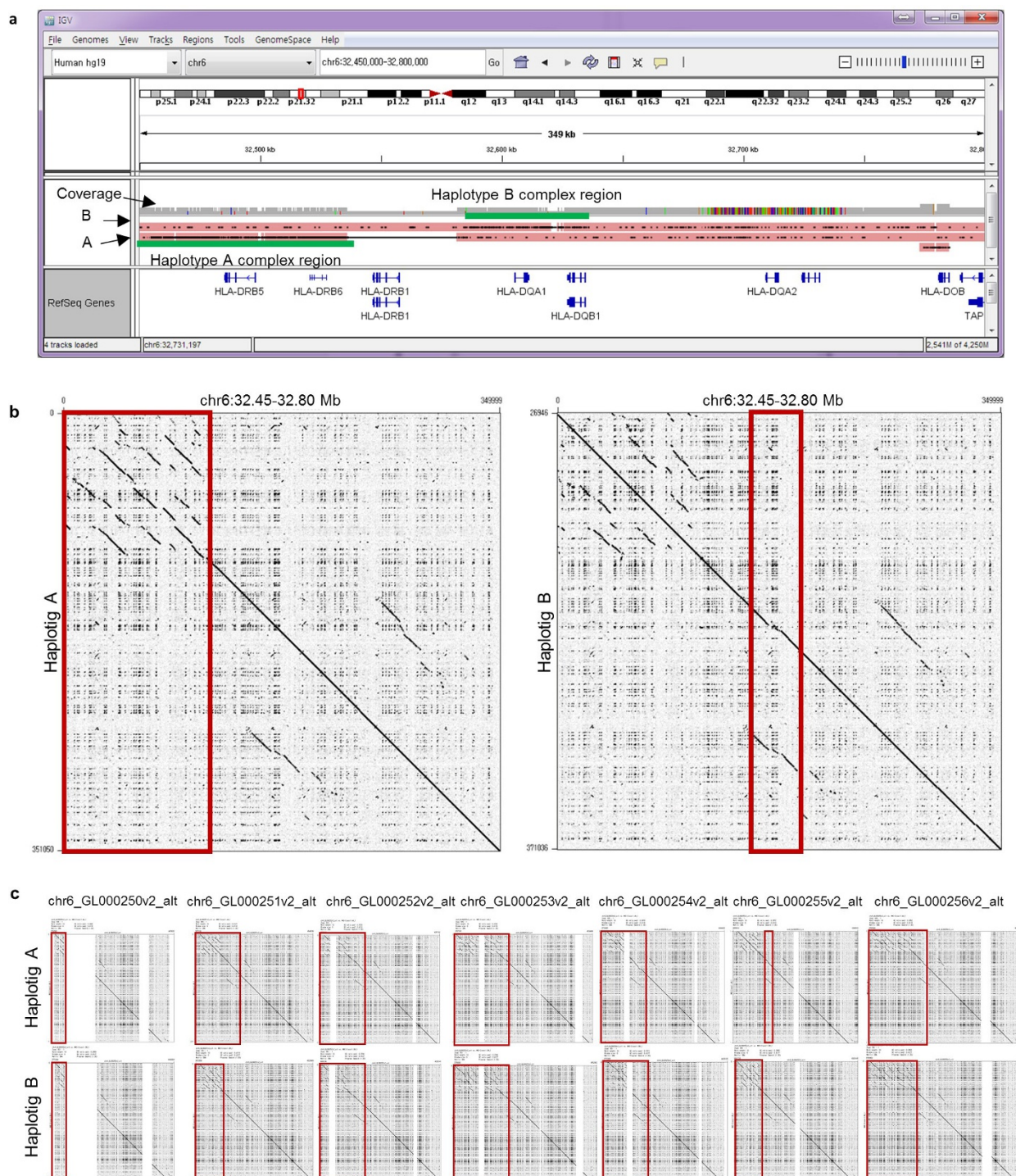
Extended Data Figure 7 | MHC class II haplotigs alignment on chromosome 6 and dot plots. a, MHC haplotigs A and B aligned on GRCh37 chr6. The complex regions shown in Fig. 3a are in green bars. **b,** Dot plot of haplotig A and B to the reference genome. The region highlighted in red is giving many SVs when aligning on the reference owing to different sequence context in haplotigs. **c,** Dot plots of haplotig A and B to the alternative loci (ALT) patches of MHC region in hg38. Haplotig A had the most similarities with chr6_GL000255v2_alt for the highlighted region in **b**. The blank vertical lines indicate 'N' bases in reference ALT sequence.

Extended Data Table 1 | Summary of *de novo* assembly and phasing statistics

Data set		No. of contigs or scaffolds	Sum (Gb)	N50 (Mb)	Longest (Mb)	Average (Mb)	L50	No. of gaps	Percent bases of gaps
Contigs	Assembly	3,128	2.87	17.7	76.5	0.92	50	0	0.00
Scaffold V1	Scaffolding	2,927	2.89	29.1	113.9	0.99	28	188	0.75
Scaffold V2	Scaffolding with Fsr-GM	2,832	2.90	44.8	113.9	1.03	21	264	1.29
Phased Block	Linked-reads	1,468	2.62	5.70	29.6	1.78	143		n/a
	Linked-reads with BAC	836	2.64	11.55	65.0	3.16	71		
Haplotig	A	3,155	2.63	2.41	11.21	0.83	328		n/a
	B	15,816	2.19	0.32	2.52	0.14	2,012		

Phased blocks are measured by sub-read coverage based on markers on each step. These subreads are locally assembled to build haplotigs. A, haplotig assembled with subreads phased as haplotype A and homozygous; B, haplotig B assembled with subreads phased as haplotype B; BAC, bacterial artificial chromosome sequenced with short reads; Fsr-GM, fragile site rescued Genome Map; n/a, not applicable.

Extended Data Table 2 | BAC clone paired-end sequence placements to AK1 and human reference

Type of Placement	GRCh37	GRCh38	AK1 v1	AK1 v2
Paired End Placements	62,076	62,091	62,038	62,022
Unique Placements	60,585	60,460	61,152	61,132
Concordant Placements	56,328	56,486	58,359	58,340
Scaffolding Placements	0	0	3,027	2,928
Discordant Placements	4,257	3,974	2,793	2,792
Discordant in Size	1,481	1,358	1,310	1,316
Discordant in Orientation	1,619	1,555	1,483	1,476
Discordant in Chromosomes	1,157	1,061	na	na
Multiple Placements	1,491	1,631	886	890
Orphan Placements	655	642	698	713
Unmapped	27	25	22	23
Mean in silico insert size (bp)	102,788	102,785	102,928	102,947
In silico insert size standard deviation (bp)	23,149	23,151	23,263	23,295
Total	62,758	62,758	62,758	62,758

The summary of BAC paired end-read alignments to the human reference and the AK1 assembly. The end sequence placements indicate that the assembly quality of AK1 V2 is comparable to that of the human reference genome.

Extended Data Table 3 | Haplotig length and reference coverage on each chromosome

	Haplotig A					Haplotig B				
	No. haplotigs	Bases (Mb)	Coverage	N50 (Kb)	Longest (Mb)	No. haplotigs	Bases (Mb)	Coverage	N50 (Kb)	Longest (Mb)
chr1	175	217.1	96%	2,218.1	7.8	1,262	182.1	81%	337.8	1.5
chr2	179	229.9	96%	2,325.1	9.3	1,403	193.8	81%	315.5	1.5
chr3	139	194.8	100%	2,282.2	8.8	1,250	156.7	80%	297.8	1.6
chr4	140	174.6	93%	2,380.3	10.7	1,084	153.0	82%	311.4	1.6
chr5	122	172.9	97%	2,296.1	11.3	1,064	143.2	81%	300.3	1.3
chr6	104	165.7	99%	2,048.4	6.7	969	136.5	82%	325.0	1.4
chr7	123	149.1	96%	2,294.4	11.8	878	125.3	81%	316.0	1.8
chr8	97	144.2	101%	2,340.6	7.4	850	121.6	85%	316.0	3.4
chr9	88	112.1	93%	2,746.8	7.2	565	96.6	80%	377.8	2.6
chr10	114	125.1	95%	2,485.2	4.5	770	106.1	81%	328.2	1.4
chr11	100	126.1	96%	2,606.4	7.9	724	106.8	81%	346.5	1.8
chr12	73	128.5	98%	2,958.9	10.9	717	107.5	82%	312.1	1.5
chr13	56	95.0	99%	2,975.3	7.4	572	79.1	83%	286.6	1.2
chr14	56	86.1	97%	2,675.1	6.6	457	75.2	85%	361.8	1.5
chr15	74	76.5	94%	2,530.2	6.1	431	63.5	78%	304.8	1.0
chr16	92	73.0	93%	2,196.8	7.3	433	61.7	78%	305.7	1.2
chr17	70	74.8	96%	1,367.1	5.4	489	61.8	79%	269.0	1.3
chr18	43	73.6	99%	3,309.4	7.7	414	64.1	86%	296.5	1.3
chr19	51	55.0	99%	1,853.4	4.4	316	48.2	86%	291.6	1.6
chr20	31	57.9	97%	3,914.5	7.6	328	49.8	84%	349.4	1.5
chr21	35	33.0	94%	1,980.9	12.5	180	29.1	83%	304.1	0.8
chr22	30	33.1	95%	2,429.5	6.2	159	29.5	85%	366.0	1.4
Total	1,992	2,598.1	97%	2,403.9	12.5	15,315	2,191.5	82%	318.7	3.4

Coverage was calculated over autosomal non-N base length of the reference (GRCh37). Haplotig A in chromosome 8 is even longer than the non-N bases of the reference, indicating both assembly and phasing achieved higher contiguity.