nature genetics
OPEN

Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome

Derek M Bickhart[1,18], Benjamin D Rosen[2,18], Sergey Koren[3,18], Brian L Sayre[4], Alex R Hastie[5], Saki Chan[5], Joyce Lee[5], Ernest T Lam[5], Ivan Liachko[6], Shawn T Sullivan[7], Joshua N Burton[6], Heather J Huson[8], John C Nystrom[8], Christy M Kelley[9], Jana L Hutchison[2], Yang Zhou[2,10], Jiajie Sun[11], Alessandra Crisà[12], F Abel Ponce de León[13], John C Schwartz[14], John A Hammond[14], Geoffrey C Waldbieser[15], Steven G Schroeder[2], George E Liu[2], Maitreya J Dunham[6], Jay Shendure[6,16], Tad S Sonstegard[17], Adam M Phillippy[3], Curtis P Van Tassell[2] & Timothy P L Smith[9]

The decrease in sequencing cost and increased sophistication of assembly algorithms for short-read platforms has resulted in a sharp increase in the number of species with genome assemblies. However, these assemblies are highly fragmented, with many gaps, ambiguities, and errors, impeding downstream applications. We demonstrate current state of the art for *de novo* assembly using the domestic goat (*Capra hircus*) based on long reads for contig formation, short reads for consensus validation, and scaffolding by optical and chromatin interaction mapping. These combined technologies produced what is, to our knowledge, the most continuous *de novo* mammalian assembly to date, with chromosome-length scaffolds and only 649 gaps. Our assembly represents a ~400-fold improvement in continuity due to properly assembled gaps, compared to the previously published *C. hircus* assembly, and better resolves repetitive structures longer than 1 kb, representing the largest repeat family and immune gene complex yet produced for an individual of a ruminant species.

A finished, accurate reference genome is essential for advanced genomic selection of productive traits and gene editing in agriculturally relevant plant and animal species[1–3]. Thus, efficient genome finishing technologies will be of immediate benefit to researchers of these organisms. Substantial progress has been made in methods for generating contigs from whole-genome shotgun (WGS) sequencing; yet finishing genomes remains a labor-intensive process that is unfeasible for most large, highly repetitive genomes. The successful production of the human reference genome assembly draft in 2001 (ref. 4) was followed by 3 years of intensive curation by 18 individual institutions[5] to produce the best available reference genome assembly for a mammalian species, of which the current version (GRCh38) contains only 832 heterochromatin-associated gaps. Although inexpensive short-read sequencing has enabled the creation of a substantial number of draft genome assemblies, they are highly fragmented because high-throughput methods for finishing were not available[6].

Repeats pose the largest challenge for reference genome assembly, and much effort has been devoted to resolving the ambiguous assembly gaps caused by repetitive DNA sequence[7]. Numerous scaffolding technologies have been developed for ordering and orienting assembly contigs[8–12], including chromosome interaction mapping (Hi-C)[13] and optical mapping[14], which provide relatively inexpensive and high-resolution scaffolding data[15–19]. Hi-C is an adaptation of the chromosome conformation capture (3C) methodology[20] that identifies long-range chromosome interactions in an unbiased fashion without *a priori* target site selection. The frequency of long-range consensus interactions decays rapidly as linear distance along a chromosome increases, allowing Hi-C data to scaffold assembled contigs to the scale of full chromosomes[15]. Optical mapping technologies observe the linear separation of small DNA motifs (often restriction enzyme

[1]Cell Wall Biology and Utilization Laboratory, ARS USDA, Madison, Wisconsin, USA. [2]Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, Maryland, USA. [3]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA. [4]Department of Biology, Virginia State University, Petersburg, Virginia, USA. [5]BioNano Genomics, San Diego, California, USA. [6]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [7]Phase Genomics, Seattle, Washington, USA. [8]Department of Animal Science, Cornell University, Ithaca, New York, USA. [9]US Meat Animal Research Center, ARS USDA, Clay Center, Nebraska, USA. [10]Shaanxi Key Laboratory of Agricultural Molecular Biology, College of Animal Science and Technology, Northwest A&F University, Yangling, China. [11]South China Agricultural University, Tianhe, Guangzhou, China. [12]Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA)–Animal Production Research Centre, Rome, Italy. [13]Department of Animal Science, University of Minnesota, St. Paul, Minnesota, USA. [14]Livestock Viral Diseases Programme, The Pirbright Institute, Woking, UK. [15]Warmwater Aquaculture Research Unit, ARS USDA, Stoneville, Mississippi, USA. [16]Howard Hughes Medical Institute, Seattle, Washington, USA. [17]Recombinetics, Inc., St. Paul, Minnesota, USA. [18]These authors contributed equally to this work. Correspondence should be addressed to A.M.P. (adam.phillippy@nih.gov), C.P.V.T. (curt.vantassell@ars.usda.gov) or T.P.L.S. (tim.smith@ars.usda.gov).

Received 2 August 2016; accepted 3 February 2017; published online 6 March 2017; doi:10.1038/ng.3802

recognition sites[19] or nickase sites[21]), which can provide sufficient contextual information to scaffold assembled contigs[22] or correct existing reference assemblies[23]. Both optical mapping[21] and Hi-C[15] yield excellent scaffold continuity metrics[15,17,18,24]. However, both methods have limited ability to scaffold small contigs in fragmented short-read assemblies[25].

Single-molecule sequencing[26] can now produce reads tens of kilobases in size, albeit with relatively high error rate. The Pacific Biosciences PacBio RSII sequencing platform achieves an average read length of 14 kb, with maximum read lengths >60 kb[27], and is routinely used to reconstruct complete bacterial genomes[28,29] and highly continuous eukaryotic genomes[27,30,31]. When maximum read length exceeds the maximum repeat size, it is theoretically possible to assemble complete mammalian chromosomes. However, the read depth required to ensure that all repeats are spanned by such reads is currently prohibitive, so mammalian assemblies will continue to comprise thousands of pieces[27,30] until average read lengths exceed ~30 kb. Currently, combinations of long-read sequencing and long-range scaffolding represent the most efficient approach to produce near-finished reference assemblies. For example, a recent study using long-read sequencing and optical mapping assembled a human genome *de novo* into 4,007 contigs and 202 scaffolds that covered the entire reference assembly[31].

Here we present a near-finished reference genome for the domestic goat (*C. hircus*) using a combination of long-read single-molecule sequencing, high-fidelity short-read sequencing, optical mapping, and Hi-C-based chromatin interaction maps. Unlike cattle, which are derived from two different subspecies[32], extant domestic goats appear to derive from a single wild ancestor, the bezoar[33]. Owing to this singular domestication event, creation of a polished reference genome for goat could enable easier identification of adaptive variants in sequence data from descendent breeds. The most recent goat assembly was generated via short-read sequencing and optical mapping and is highly fragmented[18]. Our new assembly strategy achieves superior continuity and accuracy, is cost effective compared to past finishing approaches, and provides a new standard reference for ruminant genetics.

RESULTS

De novo assembly of a *C. hircus* reference genome

We sequenced an adult male goat of the San Clemente breed with a high degree of homozygosity to minimize heterozygous alleles and simplify assembly. A combination of three technologies was applied: single-molecule real-time sequencing (PacBio RSII), paired-end sequencing (Illumina HiSeq), and Hi-C (Phase Genomics, Inc.). We also generated optical mapping (using BioNano Genomics Irys) data, but these came from an adolescent male progeny of the reference animal owing to tissue storage complications. Assembly of these complementary data types proceeded in a stepwise fashion (Online Methods), producing progressively improved assemblies (**Table 1** and **Fig. 1**). Initial assembly of the PacBio data alone resulted in a contig NG50 (the minimum length of contigs accounting for half of the haploid genome size) of 3.8 Mb. PacBio contigs were first scaffolded using optical mapping data, and the resulting scaffolds were clustered using Hi-C data into chromosome-scale scaffolds. To assess quality, the resulting assembly was validated via statistical methods and comparison to a radiation hybrid (RH) map[34] (**Supplementary Table 1**) and previous assemblies (**Supplementary Note**). To maximize accuracy of the final reference assembly, the RH map was used to correct 21 inversions (consisting of 83 scaffolds) and 4 misplacements before final gap filling and polishing[35,36]. Our final assembly, ARS1, totaled 2.92 Gb of sequence with a contig NG50 of 18.7 Mb, a scaffold NG50 of 87 Mb, and an estimated quality value (QV)[37] of 34.5 (**Table 1**, **Fig. 2** and **Supplementary Note**). After error correction and validation, ARS1 contained four major disagreements with the RH map (**Fig. 3**), which will require further investigation to confirm. Considering that ARS1 comprises just 31 scaffolds and 649 gaps covering 30 of the 31 haploid, acrocentric goat chromosomes[38] (excluding only the Y chromosome), our assembly compares favorably with the current human reference (GRCh38), which has 24 scaffolds, 169 unplaced or unlocalized scaffolds, and 832 gaps in the primary assembly[39].

Scaffolding technology comparisons

We compared initial *de novo* optical map and Hi-C scaffolds to our final validated reference assembly to evaluate the independent performance of the two scaffolding strategies. The optical map consisted of 2,944 scaffolds with an NG50 of 1.487 Mb. It is likely that optical map fragment sizes (**Supplementary Fig. 1**) were limited by double-strand breaks caused by close proximity of Nt.BsqI sites on opposing DNA strands, as reported previously[21]. Optical map scaffolding of PacBio contigs produced an assembly of 333 scaffolds, containing 90.89% of the final ARS1 assembly length with a scaffold NG50 of 20.623 Mb, and identified 36 misassemblies in the PacBio contigs. This twofold increase in NG50 value over the individual technologies (**Table 1**) is likely due to the complementary nature of their error profiles; the long PacBio reads span shorter, low-complexity repeats, whereas the optical map spans larger segmental duplications. In comparison, scaffolding of PacBio contigs with Hi-C data yielded 31 scaffolds containing 87.9% of the total assembly length (**Table 1**, **Supplementary Fig. 2** and **Supplementary Table 2**). These scaffolds had an NG50 four times larger than that of the scaffolds generated by optical mapping, but their rate of misoriented contigs was high in comparison to the RH map[34] (**Supplementary Note**). Analysis of the misoriented contigs revealed that orientation error was correlated

Table 1 Assembly statistics

Assembly[a]	Contigs[b]	Scaffolds	Unplaced contigs[c]	Degenerate contigs[d]	Contig NG50 (Mb)[e]	Scaffold NG50 (Mb)[e,f]	Assembly size (Gb)	Assembly in scaffolds (%)
PacBio	3,074	–	–	30,693	3.795	–	2.914	N/A
Optical Map	–	2,944	–	–	–	1.487	2.748	N/A
PacBio + Optical Map	1,109	333	1,242	30,693	10.197	20.623	2.910	90.89
PacBio + Hi-C	2,115	31	959	30,693	3.795	88.799	2.910	87.97
PacBio + Optical Map + Hi-C	1,780	31	571	30,693	10.197	87.347	2.910	89.05
ARS1	680	31	654	29,315	18.702	87.277	2.924	88.32

[a]Assemblies are listed in order of inclusion of scaffolding technologies toward the final assembly (ARS1), with the original contigs (PacBio) scaffolded using different technologies (Optical Map and Hi-C). Because the optical map program (Irys Scaffold) generates an assembly from the consensus of labeled DNA molecules, we have included scaffold statistics from these data (optical map) for comparison. [b]The number of continuous stretches of sequence within the scaffold without gaps >3 bases in length of at least 10 bases. [c]Unplaced contigs are defined as input contigs or scaffolds that were not placed by the optical map or Hi-C in a scaffold were excluded from the scaffold counts. [d]Degenerate contigs were assembled unitigs that had less than 50 PacBio reads supporting their assembly (**Supplementary Note**). Differences in degenerate contig counts in the final ARS1 assembly are due to PBJelly merging of degenerate contigs (538 contigs) or removal due to no supporting PacBio read alignments (840). [e]All NG50 values are based on the ARS1 assembly size (2.924 Gb). [f]No scaffolds were generated for the PacBio entry.

with the density of Hi-C restriction sites in the contig (**Supplementary Table 3**), which might be improved by choosing restriction enzymes with shorter recognition sites (or DNase Hi-C)[40] to improve Hi-C link density and reduce the associated orientation error rate. Ultimately, we found that sequential scaffolding with optical mapping data followed by Hi-C data yielded an assembly with the highest continuity and best agreement with the RH map (**Fig. 1**). Thus, the final ARS1 assembly was based on this approach and the remaining inversions found in comparisons to the RH map were corrected manually before final gap filling and polishing.

Assembly benchmarking and comparison to reference

The goat CHIR_1.0 reference assembly[18] was generated from paired-end short reads using the SOAPdenovo2 assembler, a restriction-enzyme-based optical map, and cross-species scaffold alignments to the *Bos taurus* UMD3.1 reference assembly[41]. The CHIR_2.0 assembly (GenBank GCA_000317765.2) is a recent improvement to the CHIR_1.0 assembly that used the goat radiation hybrid map data for scaffolding and probably included additional curation but has not yet been described. Paired-end read sequences used to create the Black Yunan goat CHIR_1.0 reference assembly[18] were aligned to CHIR_1.0, CHIR_2.0, and our ARS1 assembly for a reference-free measure of structural correctness[42–44] (**Supplementary Note**). These alignments confirmed that CHIR_2.0 is a general improvement over CHIR_1.0, with fewer putative deletions (2,735 versus 10,256) and duplications (115 versus 290); however, CHIR_2.0 also contains 50-fold more putative inversions than CHIR_1.0 (215 versus 4) (**Supplementary Table 4**). Our ARS1 assembly is a further improvement over CHIR_2.0, with 4-fold fewer deletions and 50-fold fewer inversions identified. This is particularly notable given that the Black Yunan data were not used for constructing ARS1, yet our assembly is more consistent with the Black Yunan paired-end data than the CHIR_1.0 and CHIR_2.0 assemblies themselves. We assessed large-scale structural continuity of each assembly by aligning fosmid end sequence and identifying structural variants (Online Methods and **Supplementary Table 5**). ARS1 had half the number of trans-scaffold discrepancies ('break end'(BND) variants: 456) of CHIR_2.0 (840) and had 13 fewer assembly errors per 100 Mb. This independent validation suggests that ARS1 corrects numerous errors present in CHIR_2.0 (**Fig. 2**).

We also assessed the quantity and size of gaps in each respective assembly (**Supplementary Table 6**). The CHIR_2.0 reference filled 62.4% of CHIR_1.0 gap sequences (160,299 gaps filled), whereas our assembly filled 94.3% of all CHIR_1.0 gaps (242,268 gaps filled). The remaining CHIR_1.0 gaps (13,853) had flanking sequence that mapped to two separate chromosomes in our assembly, indicating potential false gaps due to errors in the CHIR_1.0 assembly. WGS sequence alignments from our San Clemente reference animal as well as alignments of gap fill regions from CHIR_2.0 agreed with our assembly in closed gap locations (Online Methods), revealing 200,624 CHIR_1.0 gaps (77.02% of total) confirmed as closed in ARS1. Of the remaining 59,850 CHIR_1.0 gaps that were not confirmed as closed in ARS1, 52 coincided with gaps in ARS1, 568 were predicted to be filled by greater than 10 kb of sequence, and 23 did not have flanking sequence that could be mapped to the ARS1 assembly. Because gaps coinciding with ARS1 gaps are currently ambiguous, it is difficult to ascertain the true status of these remaining regions. Fosmid end structural variant calls (**Supplementary Table 5**) intersected 14 of ARS1 gap regions, suggesting that there are structural discrepancies or assembly errors that contribute to the unknown gaps in ARS1. In total, our assembly contains 649 sequence gaps (larger than 3 bp) in the chromosomal scaffolds split among gaps of known (515 inferred from optical mapping distances) and unknown (134 Hi-C scaffold joining) sizes. Compared to CHIR_2.0, ARS1 has 1,000-fold fewer ambiguous bases and improves even the core gene annotation over the short-read assembly by receiving a 2-point higher BUSCO score[45] (82% versus 80%, respectively).

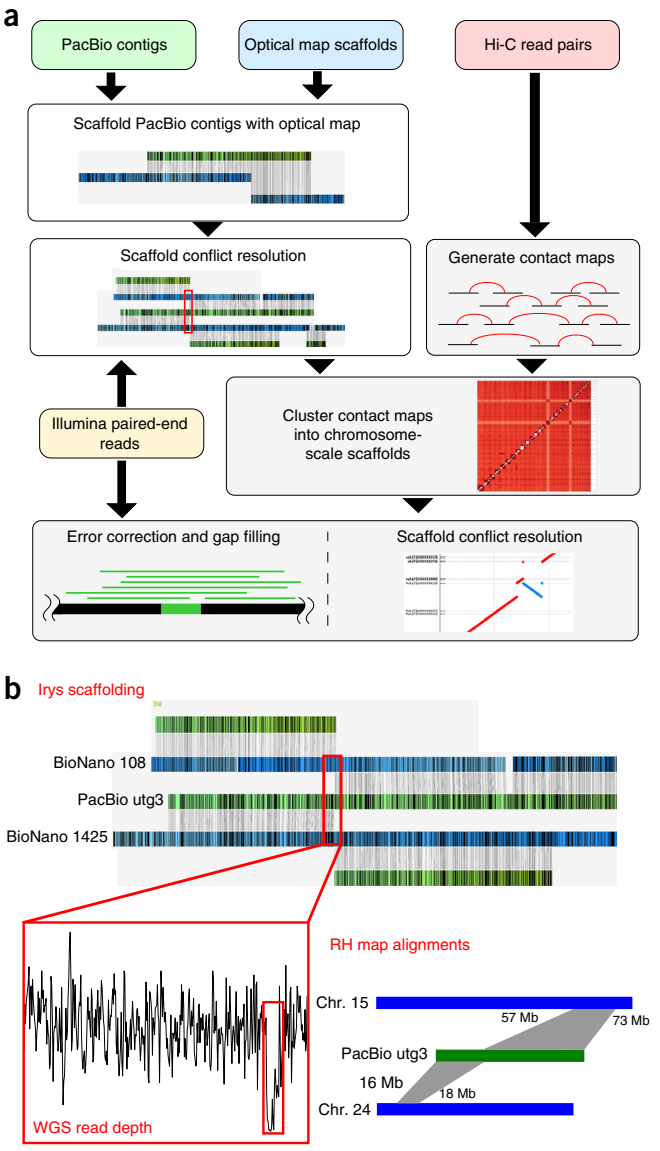

Figure 1 Assembly schema for producing chromosome-length scaffolds. (**a**) Four sets of sequencing data (long-read WGS, Hi-C, optical mapping, and short-read WGS) were produced to generate the goat reference genome. A tiered scaffolding approach using optical mapping data followed by Hi-C proximity-guided assembly produced the highest-quality genome assembly. (**b**) An example from the initial optical mapping data set. To correct misassemblies resulting from contig or scaffold errors, a consensus approach was used. A scaffold fork was identified on contig 3 (91 Mb long) from the optical mapping data. Mapping of short-read WGS data signature showed a misassembly near the thirteenth megabase of the contig, so it was split at this region. Subsequent analysis based on the RH map confirmed this split.

Improved genetic marker tools and functional annotation

We quantified the benefit of our approach over short-read assembly methods with respect to genome annotation and downstream functional

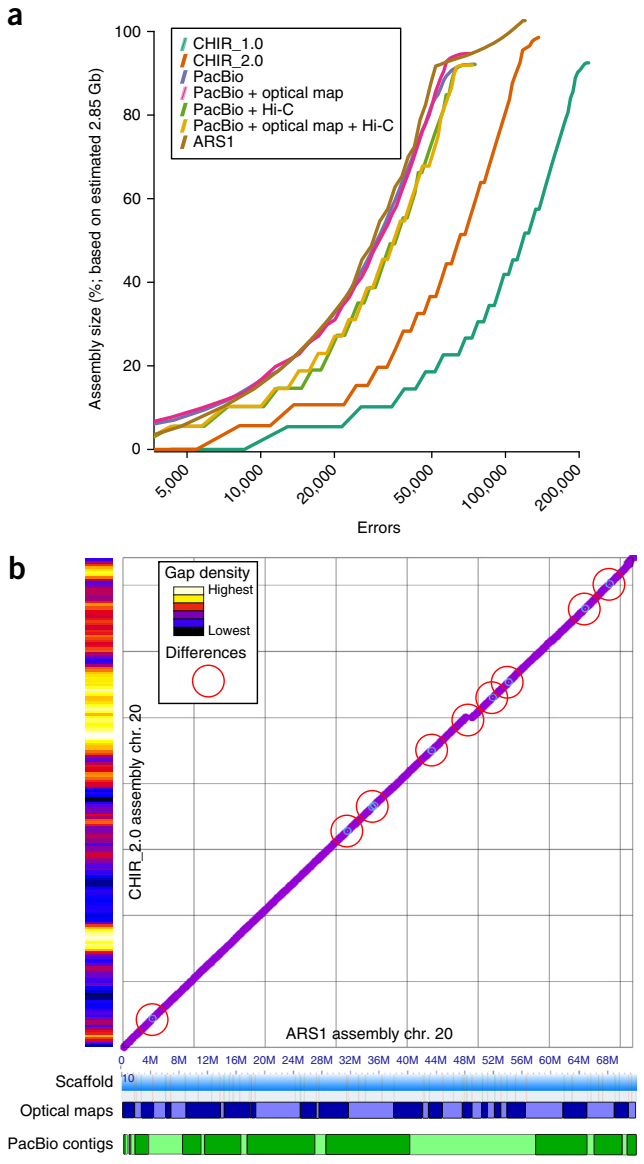

Figure 2 Assembly benchmarking comparisons reveal high degree of assembly completion. (**a**) Feature response curves showing the error rate as a function of the number of bases in each assembly (CHIR_1.0, CHIR_2.0, and ARS1) and each scaffold test (intermediary assemblies using a combination of Hi-C and BioNano scaffolding). (**b**) Comparison plots of chromosome 20 sequence between the ARS1 and CHIR_2.0 assemblies reveal several small inversions (light blue circles) and a small insertion of sequence (break in continuity) in the ARS1 assembly. Red circles highlight inversions and the insertion of sequence in our assembly. ARS1 optical map scaffolds and PacBio contigs are represented below the *x* axis.

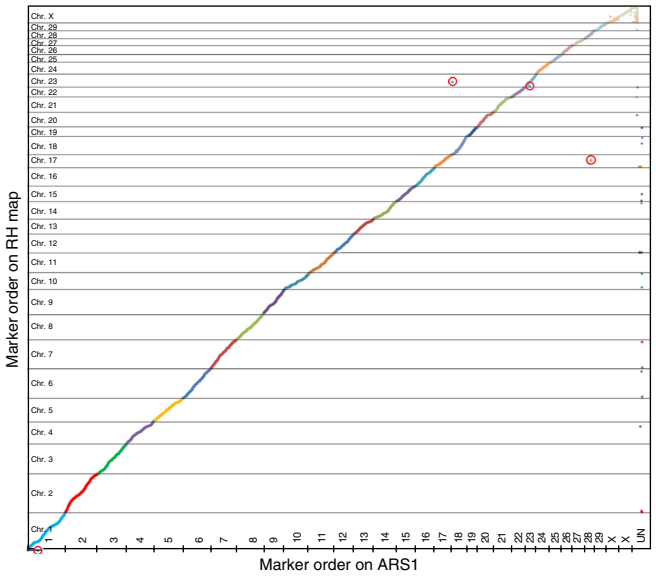

Figure 3 RH probe map shows excellent assembly continuity. ARS1 RH probe mapping locations were plotted against the RH map order. Each ARS1 scaffold corresponds to an RH map chromosome, with the exception of X, which is composed of two scaffolds. Red circles highlight two intrachromosomal (on chr. 1 and chr. 23) and two interchromosomal misassemblies (on chr. 18 and chr. 17) in ARS1 that were difficult to resolve.

analysis. Chromosome-scale continuity of the ARS1 assembly was found to have appreciable positive impact on genetic marker order for the existing *C. hircus* 52K SNP chip[3] (**Supplementary Table 7**). Of the 1,723 SNP probes currently mapped to the unplaced contigs of the CHIR_2.0 assembly, we identified chromosome locations for 1,552 unplaced markers (90.0% of 1,723 unplaced) and identified 26 markers with ambiguous mapping locations (1.8% of 1,466 low-call rate markers)[3]. This finding suggests that the latter markers were targeting repeat sequences and may explain why their call rate was poor.

After annotation, we found 3,495 newly annotated gene models (Online Methods) that contained at least one gap in the CHIR_2.0 assembly that was filled by our assembly (**Supplementary Table 6**). We also identified 1,926 predicted exons that contained gaps in CHIR_1.0 and CHIR_2.0 but were resolved by our assembly (**Fig. 4a**), probably owing to an improvement in resolution of repetitive content (**Fig. 4b**). Notably, annotation of repetitive immune-associated gene regions revealed that complete complements of the genes encoding leukocyte receptor complex (LRC) and natural killer cell complex (NKC) were contained within single autosomal scaffolds in our assembly (**Fig. 5**). These regions are particularly difficult to assemble with short-read technologies because they are highly polymorphic and repetitive[46], and their gene content is largely species specific. We think the successful assembly and annotation of these regions in ARS1 is an important achievement (**Supplementary Note** and **Supplementary Figs. 3–5**).

Structural elements and karyotype

The combination of technologies used for ARS1 substantially improves on repeat resolution compared to previous assembly approaches, including both short-read and Sanger sequencing projects[41,47]. Large fractions of the Y chromosome and heterochromatin regions were assembled, whereas these are typically absent from *de novo* assembly efforts. For example, the presence of >5 bp of telomeric sequence on six autosomes indicates that scaffolds have reached one end of the acrocentric chromosomes. Using previously determined centromeric repeat sequence for goat[48], we identified 15 chromosome scaffolds that included centromeric repeats >2 kb in length (Online Methods), suggesting inclusion of the centromeric ends. Seven chromosomes (1, 6, 12, 13, 22, 26, and 29) had centromeric repeat sequence alignments that were >8 kb in length. Chromosomes 19 and 23 had centromere and telomere repeats on opposite ends, consistent with complete chromosome-wide assembly. Two scaffolds (corresponding to chromosomes 13 and 28) had centromeric repeats 3 Mb from the end, suggesting that the ARS1 assembly includes the elusive p arm

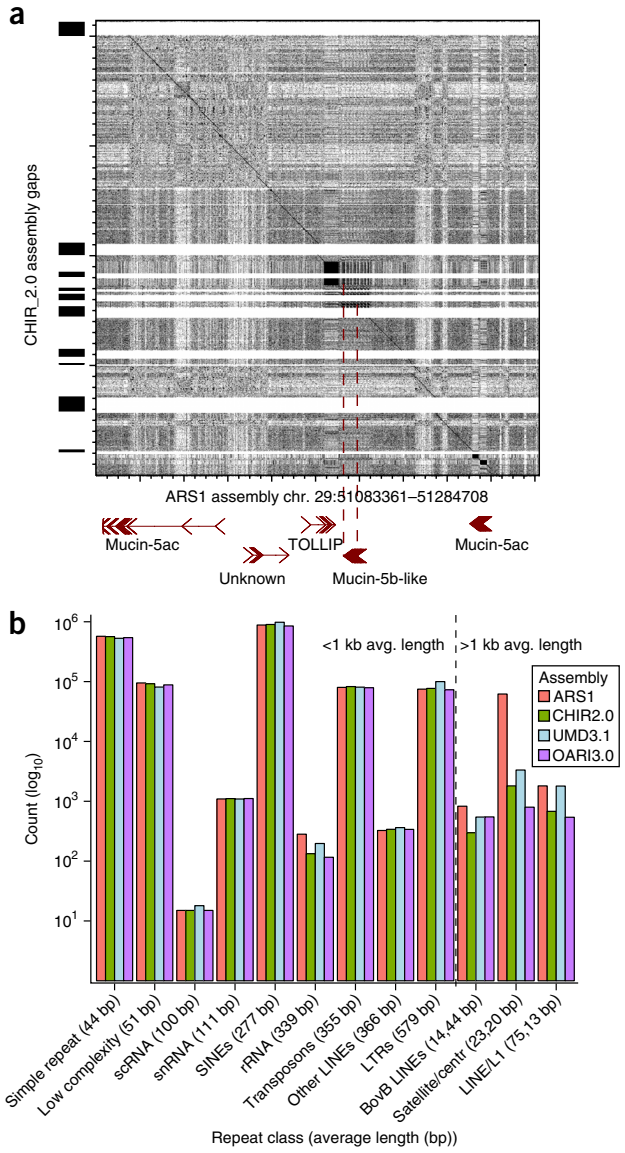

Figure 4 Long-read assembly with complementary scaffolding resolves gap regions and long repeats that cause problems for short-read reference annotation. (**a**) A region of the mucin gene cluster was resolved by long-read assembly, resulting in a complete gene model for *LOC107345534* (mucin-5B-like). (**b**) Counts of repetitive elements that had greater than 75% sequence length and greater than 60% identity with RepBase database entries for ruminant lineages.

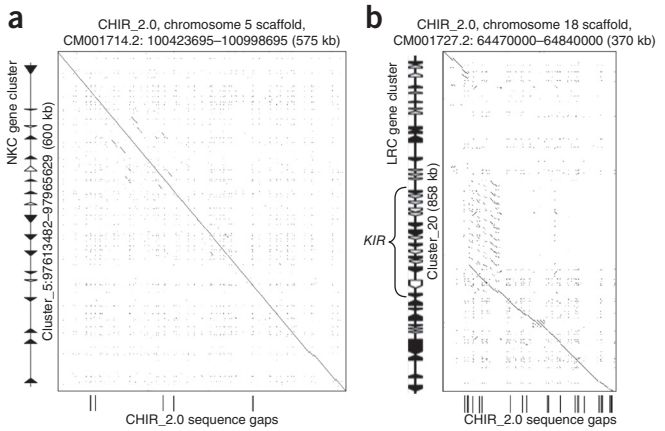

Figure 5 Comparative alignment of resolved immune gene clusters in the domestic goat. (**a**) A region of the natural killer cell (NKC) gene cluster contained several gaps in the CHIR_2.0 reference genome (*x* axis) but was present on a single contig on the ARS1 assembly (*y* axis). (**b**) The leukocyte receptor complex (LRC) locus was poorly represented in CHIR_2.0 (*x* axis) and was missing ~500 kb of sequence that is present in ARS1 (*y* axis).

of these acrocentric chromosomes (Online Methods). Additionally, closer examination of the optical maps revealed 34 maps containing large tandem and interspersed repetitive nickase motifs, with a cumulative size of 4 Mb, that did not align to the long-read contigs (**Supplementary Table 8**). Because these repetitive maps also did not align to any prior *C. hircus* assembly, they may represent constitutive heterochromatin that could not be assembled using other technologies. We identified 105 additional repetitive patterns >12 kb in the optical map that were represented in ARS1, distributed among all Hi-C chromosome scaffolds except chromosomes 9 and 10. Finer-scale repeat identification using the RepeatMasker algorithm confirmed that the larger classes of repetitive elements (>1 kb) were resolved in ARS1 (**Fig. 4b**), and 66% more BovB LINE repeats were assembled to at least 75% of the repeat length than in CHIR_2.0. Notably, 43.6% of the CHIR_2.0 gaps that ARS1 successfully closed coincided with BovB repeats >3.5 kb in length (**Supplementary Fig. 6** and **Supplementary Table 9**). Comparison of fosmid end sequence data to repetitive sequence identified only five structural variants (two predicted duplications, two predicted deletions and one inversion) that intersected with our larger repetitive regions, including the predicted centromeric region on chromosome 10 (**Supplementary Note**), suggesting that at least five large repeats (5/30,347 repeats >1 kb, or 0.016% of identified repeats) in ARS1 may be misassembled.

The final ARS1 assembly contained two scaffolds that mapped to two different—but continuous—regions of the X chromosome, representing 85.9% of the expected chromosome size (assuming a size of 150 Mb)[38]. Self-hit alignment filtering, and cross-species alignment to existing Y chromosome scaffolds in cattle, identified 10 Mb of sequence that may have originated from the *C. hircus* Y chromosome, ~50% of the estimated size[49] (**Supplementary Note** and **Supplementary Table 10**). Alignments of X-degenerate Y genes[50] and *B. taurus* Y genes to these scaffolds confirmed their association with the Y chromosome, identifying 16% and 84% of our self-hit filtered contig list, respectively, with several contigs containing both sets of alignments. Both the heterochromatic nature of the Y chromosome and the ambiguous placement of the pseudo-autosomal region on the X or Y chromosome (the last portion of our X chromosome and unplaced scaffolds 8, 12, 119, and 186) precluded generation of chromosome-scale scaffolds for the male sex chromosome.

DISCUSSION

The advent of long-read sequencing has dramatically improved the average and N50 contig lengths of mammalian genome assemblies[27,31], but complex genomic regions still interfere with the generation of complete, single-contig chromosomes[31]. Attempts to fill gaps in existing short-read assemblies with low-coverage long reads fail to close many gaps that could otherwise be closed with higher coverage[51], as shown by the ~41,000 gaps remaining in the *Ovis aries* Oar_v4.0 assembly (ENA GCA_000298735.2) and the ~35,000 gaps in the *B. taurus* Btau_5.0.1 assembly (ENA GCA_000003205.6). Complex genomic regions have even higher impact for genomes that are polyploid or have historical whole-genome duplications. Increasing coverage means that a more very long reads from the top

tail of the read-length distribution are collected, and this helps resolve large repetitive regions. Thus, higher coverages of long reads tend to provide superior results to gap-filled short-read assemblies, as demonstrated by the few gaps remaining in ARS1. However, current long-read technologies still fall short of regularly producing completely assembled chromosomes, so reliable and affordable scaffolding technologies remain vitally important for generating high-quality finished reference genome assemblies. In this study we assessed the utility of both optical and chromatin interaction mapping, showing that they are complementary and particularly useful in combination with long-read assemblies. Stepwise combination of these methods leveraged their unique benefits to generate a final assembly.

Optical mapping had fewer conflicts with the initial contigs and provided higher resolution, so the resulting scaffolds were easier to validate than the Hi-C scaffolds. However, optical mapping was insufficient to generate full chromosome-scale scaffolds, with the notable exception of the single scaffold spanning goat chromosome 20 (**Fig. 2b**). The primary limitation of the goat optical map appears to be double-strand breaks caused by neighboring nickase sites on opposite strands, which breaks the map assembly owing to a lack of spanning fragments[21]. Optical map scaffolding generated only three confirmed assembly errors (3/333, or 0.9% of scaffolds), two of which were difficult to detect without the use of the RH map. Scaffolding with Hi-C enabled accurate assignment of contigs to their respective chromosome groups, as supported by our RH map data, 99.8% of the time; however, there were 21 confirmed order and orientation errors affecting 83 scaffolds (83/1,533; 5.41%). Misorientation by Hi-C could be reduced with longer input contigs, higher numbers of orienting restriction sites, or selection of a restriction enzyme with a higher frequency of recognition sites. Contigs and scaffolds with low orientation quality scores were frequently associated with orientation mistakes in the Hi-C scaffolds (Pearson's $r = 0.49$) (**Supplementary Table 3**), suggesting that more frequent cutting may provide higher fidelity.

Optical mapping and Hi-C scaffolding had distinct error profiles. The Hi-C method was more likely to invert smaller contigs in final scaffolds, whereas the optical mapping method was more likely to leave contig errors uncorrected owing to insufficient optical map coverage. Both scaffolding methods were sensitive to the quality of the input sequence data, evident from the improvement of Hi-C scaffolding after optical scaffolding (**Table 1**) and the large relative improvement of our optical map scaffold NG50 compared to CHIR_1.0, which used optical mapping in combination with short-read data[18]. Despite these limitations, we achieved the reconstruction of 29 vertebrate autosomes into single scaffolds with a minimal number of gaps and without manual finishing (649 total gaps; 417 gaps in autosomes alone, excluding the starts and ends of chromosome scaffolds).

Mammalian genome references have generally been produced from female animals to improve coverage of the X chromosome, leaving assembly of the Y chromosome to separate, targeted projects[52,53]. Despite using a male animal, the ARS1 assembly has better X-chromosome continuity than the short-read assemblies from a female goat and produced some Y-associated scaffolds. Hi-C scaffolding was successful at clustering sex-chromosome contigs but was unable to scaffold the Y chromosome or segregate X and Y chromosome contigs into singular distinctive clusters. Optical mapping also encountered difficulty in generating Y chromosome scaffolds, generating 16 scaffolds that contained 50.2% of the putative Y chromosome sequence in our assembly. Much of the Y sequence is constitutive heterochromatin[38], which makes the generation of large optical maps and Hi-C fragments difficult.

Validation of the combined PacBio, optical map, and Hi-C assembly using the RH map demonstrated that there are limitations to the approach despite its tremendous improvement in continuity. There were 6.1% of scaffolded scaffolds, spanning 422.1 Mb (14.4%) of the assembly, that appeared to be misassembled by the two scaffolding technologies before application of RH map data. The most common problem (83 of 94 discrepancies among 1,553 scaffolds) was misorientation of contigs within scaffolds. The recommended improvements in Hi-C library preparation and optical map generation suggested here, as well as the refinement of scaffolding algorithms, could further reduce this error in future projects. Additionally, ARS1 is a haplotype-mixed representation of a diploid animal. Haplotype phasing is possible using single-molecule[54] and Hi-C[55] technologies, so a future aim is to generate a phased reference assembly.

The proposed assembly approach still has difficulty with constitutive heterochromatin, including most of the centromeres and telomeres, as well as large tandem repeats, such as the nucleolar organizer regions. Long-read contigs, optical maps, or Hi-C interaction signals cannot accurately model these features for inclusion in the assembly, and they remain unresolved even in the human reference genome, which has undergone a decade of manual finishing. Although assembly methods that can fully resolve heterochromatin regions are under development, these features are likely to remain unresolved unless sequence read lengths increase in size to routinely span them. However, ARS1 shows marked improvement in resolving the full structure of large repetitive elements, such as BovB retrotransposons and centromeric repeats (**Fig. 4b**). This increased resolution will enable future, pan-ruminant analysis of these repeat classes, which may lead to further insight into the evolution of ruminant chromosome structure.

The methods presented in this study have generated chromosome-scale scaffolds, reducing the cost of genome finishing. The tiered approach to scaffolding highly continuous single-molecule contigs obviated the need for expensive cytometry or BAC-walking experiments for chromosome placement. We estimate a current project cost of about $100,000 to complete a similar genome assembly using current RSII sequencing and the two scaffolding platforms used here. This cost is approximately three times greater than that of a short-read assembly scaffolded in a similar fashion, but the method comes with a tremendous gain in continuity and quality. The cost to achieve similar quality via manual finishing of a short-read assembly would be much higher. Moreover, advances in single-molecule sequencing, including an updated single-molecule real-time platform and alternative nanopore-based platforms, will continue to decrease this cost. As shown by the completeness of our assembly and the improvements in gene model continuity, we expect that these methods will enable the scaling of *de novo* genome assembly to large numbers of vertebrate species without requiring major sacrifices in quality.

URLs. Biowulf, https://hpc.nih.gov/systems/; Quiver FAQ, https://github.com/PacificBiosciences/GenomicConsensus/blob/master/doc/FAQ.rst; PacBio chemistry FAQ, https://github.com/PacificBiosciences/GenomicConsensus/blob/master/doc/FAQ.rst; Sheep Genome (Oarv3.1), http://www.livestockgenomics.csiro.au/sheep/oar3.1.php; RepeatMasker, http://www.repeatmasker.org/

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This study used the computational resources of the Biowulf system at the National Institutes of Health (NIH). We thank R. Lee for technical assistance. This project was supported by the US Agency for International Development Feed the Future program, the Norman Borlaug Commemorative Research Initiative, and the Livestock Improvement Program. This work was also supported in part by Agriculture and Food Research Initiative (AFRI) competitive grant 2011-67015-30183 and 2015-67015-22970 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome Program. D.M.B., B.D.R., S.G.S., and C.P.V.T. were supported by USDA CRIS project number 8042-31000-104-00. C.M.K. and T.P.L.S. were supported by USDA CRIS project number 3040-31320-012-00. G.C.W. was supported by USDA CRIS project number 6402-31000-006-00D. J. Shendure was supported in part by NIH grant R01HG006283. M.J.D. and I.L. were supported in part by NIH grant P41 GM103533. M.J.D. is a Senior Fellow in the Genetic Networks program at the Canadian Institute for Advanced Research and a Rita Allen Foundation Scholar. I.L. is supported by the UW Commercialization Gap Fund and Commercialization Fellows Program. F.A.P.d.L. was supported by MN Experiment Station Project MIN-16-103. J.C.S. and J.A.H. were funded by a UK Biotechnology and Biological Sciences Research Council Institute Strategic Program on Livestock Viral Diseases award to the Pirbright Institute. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. S.K. and A.M.P. were supported by the NIH Intramural Research Program of the National Human Genome Research Institute.

AUTHOR CONTRIBUTIONS

D.M.B., B.D.R., S.K., T.S.S., G.E.L., J. Shendure, A.M.P., C.P.V.T., and T.P.L.S. planned and coordinated the study and wrote the manuscript. T.P.L.S. and C.M.K. performed long-read sequencing and assisted with downstream analysis. S.K. and A.M.P. performed the initial long-read assembly. A.R.H., S.C., J.L., and E.T.L. performed optical mapping and provided technical support related to the data. I.L., S.T.S., J.N.B., M.J.D., and J. Shendure designed the Hi-C experiments, produced assembly scaffolds from the data, and provided technical support. D.M.B. and S.K. polished the final reference assembly. J.A.H. and J.C.S. provided manual annotation of the immune gene clusters. F.A.P.d.L. provided manual annotation of the Y chromosome genes and contigs. B.L.S. and J. Sun extracted RNA-seq biopsies and ran RNA-seq experiments, respectively. B.L.S., J.L.H., Y.Z., J. Sun, H.J.H., J.C.N., G.C.W., S.G.S., and A.C. performed downstream analysis of the data and assisted in the generation of additional files for the manuscript. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Matukumalli, L.K. *et al.* Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**, e5350 (2009).
2. Romay, M.C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* **14**, R55 (2013).
3. Tosser-Klopp, G. *et al.* Design and characterization of a 52K SNP chip for goats. *PLoS One* **9**, e86227 (2014).
4. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
6. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
7. Phillippy, A.M., Schatz, M.C. & Pop, M. Genome assembly forensics: finding the elusive misassembly. *Genome Biol.* **9**, R55 (2008).
8. Fleischmann, R.D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
9. Myers, E.W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
10. Pop, M., Kosack, D.S. & Salzberg, S.L. Hierarchical scaffolding with Bambus. *Genome Res.* **14**, 149–159 (2004).
11. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).
12. McCoy, R.C. *et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**, e106689 (2014).
13. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
14. Schwartz, D.C. *et al.* Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114 (1993).
15. Burton, J.N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
16. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
17. Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
18. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–141 (2013).
19. Nagarajan, N., Read, T.D. & Pop, M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **24**, 1229–1235 (2008).
20. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
21. Hastie, A.R. *et al.* Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome. *PLoS One* **8**, e55864 (2013).
22. Riley, M.C., Kirkup, B.C., Johnson, J.D., Lesho, E.P. & Ockenhouse, C.F. Rapid whole genome optical mapping of *Plasmodium falciparum*. *Malar. J.* **10**, 252 (2011).
23. Zhou, J., Lemos, B., Dopman, E.B. & Hartl, D.L. Copy-number variation: the balance between gene dosage and expression in *Drosophila melanogaster*. *Genome Biol. Evol.* **3**, 1014–1024 (2011).
24. Zhang, G. *et al.* Comparative genomic data of the Avian Phylogenomics Project. *Gigascience* **3**, 26 (2014).
25. 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).
26. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
27. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
28. Chin, C.-S. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
29. Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101 (2013).
30. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
31. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
32. Elsik, C.G., Tellam, R.L. & Worley, K.C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–528 (2009).
33. Naderi, S. *et al.* The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci. USA* **105**, 17659–17664 (2008).
34. Du, X.Y. *et al.* A whole-genome radiation hybrid panel for goat. *Small Rumin. Res.* **105**, 114–116 (2012).
35. English, A.C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).
36. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
37. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
38. Iannuzzi, L. & Di Meo, G.P. Chromosomal evolution in bovids: a comparison of cattle, sheep and goat G- and R-banded chromosomes and cytogenetic divergences among cattle, goat and river buffalo sex chromosomes. *Chromosome Res.* **3**, 291–299 (1995).
39. Schneider, V.A. *et al.* Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. Preprint at *bioRxiv* http://dx.doi.org/10.1101/072116 (2016).
40. Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the *cis*-regulatory landscape of human lincRNA genes. *Nat. Methods* **12**, 71–78 (2015).
41. Zimin, A.V. *et al.* A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* **10**, R42 (2009).
42. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).

43. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
44. Vezzi, F., Narzisi, G. & Mishra, B. Reevaluating assembly evaluations with feature response curves: GAGE and Assemblathons. *PLoS One* **7**, e52210 (2012).
45. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
46. Sanderson, N.D. *et al.* Definition of the cattle killer cell Ig-like receptor gene family: comparison with aurochs and human counterparts. *J. Immunol.* **1950**, 6016–6030 (2014).
47. International Sheep Genomics Consortium. The sheep genome reference sequence: a work in progress. *Anim. Genet.* **41**, 449–453 (2010).
48. Melters, D.P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
49. Hansen, K.M. Q-band karyotype of the goat (*Capra hircus*) and the relation between goat and bovine Q-bands. *Hereditas* **75**, 119–129 (1973).
50. Pérez-Pardal, L. *et al.* Multiple paternal origins of domestic cattle revealed by Y-specific interspersed multilocus microsatellites. *Heredity* **105**, 511–519 (2010).
51. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
52. Vanneste, K., Baele, G., Maere, S. & de Peer, Y.V. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
53. Tomaszkiewicz, M. *et al.* A time- and cost-effective strategy to sequence mammalian Y chromosomes: an application to the *de novo* assembly of gorilla Y. *Genome Res.* **26**, 530–540 (2016).
54. Chin, C.-S. Phased diploid genome assembly with single molecule real-time sequencing. *Nat Methods.* **12**, 1050–1054 (2016).
55. Selvaraj, S.R., Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).

ONLINE METHODS

Animals. All animal work was approved by the Virginia State University Institutional Animal Care and Use Committee. Research was conducted under an IACUC-approved protocol in compliance with the Animal Welfare Act, PHS Policy, and other federal statutes and regulations relating to animals and experiments involving animals. The facility where this research was conducted is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International, and adheres to principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 2011.

Reference individual selection. A DNA panel composed of 96 US goats from 6 breeds (35 Boer, 11 Kiko, 12 LaMancha, 15 Myotonic, 3 San Clemente, and 20 Spanish) was assembled to identify the most homozygous individual, to minimize the number of scaffold conflicts due to heterozygous genomic regions[56]. Genotypes were generated using Illumina's Caprine53K SNP beadchip processed through Genome Studio (Illumina, Inc.). The degrees of homozygosity of individuals were determined by raw counts of homozygous markers on the genotyping chip[57]. Individuals were ranked by their counts of homozygous markers, and the individual with the highest count was selected as the reference animal. An adult male of the San Clemente goat breed with 46.02% SNP-distance homozygosity (FROH) was selected from this survey as the reference animal.

Genome sequencing, assembly, and scaffolding. Libraries for SMRT sequencing were constructed as described previously[31] using DNA derived from the blood of the reference animal. We generated 465 SMRT cells using the following SMRT cell chemistry versions: P5-C3 (311 cells), P4-C2 (142 cells), and XL-C2 (12 cells) (Pacific Biosciences). A total of 194 Gb (69-fold) of subread bases with a mean read length of 5,110 bp were generated.

The Celera Assembler PacBio Corrected Reads (CA PBcR) pipeline[30] was used for assembly. Celera Assembler v8.2 was run with sensitive parameters specified by Berlin *et al.*[30], who used the MinHash Alignment Process (MHAP) to overlap the PacBio reads to themselves and PBDAGCON[28] to generate consensus for the corrected sequences. The PBcR pipeline generated 7.4 million error-corrected reads (~38 Gb; 5.1 kb average length). The error-corrected reads were in turn assembled into 3,074 contigs with an NG50 of 3.795 Mb and a total length of 2.63 Gb and 30,693 degenerate contigs—contigs with <50 supporting PacBio reads—with a total length of 288.361 Mb. Initial polishing was performed with Quiver[28] using the P5-C3 data only. The degenerate contigs (representing 9.90% of the 2.914-Gb assembled length) were excluded from scaffolding by optical maps and Hi-C and incorporated into ARS1 as unplaced contigs. Subsequent repetitive analysis revealed that 84.1% (25,821/30,693) of degenerate contigs were fully repetitive (>75% length comprised of repeats) with 94.9% (24,500/25,821) of these contigs containing a portion of centromeric or telomeric satellite sequence. The remainder were probably fragments of alternative haplotypes constituting copy number variants and other structural variants.

Scaffolding of the contigs with optical mapping was performed using the Irys optical mapping technology (BioNano Genomics). DNA of sufficient quality was unavailable from the animal sequenced owing to its accidental death, so we extracted DNA from a male offspring of the original animal. Purified DNA was embedded in a thin agarose layer and was labeled and counterstained following the IrysPrep Reagent Kit protocol (BioNano Genomics) as in Hastie *et al.*[21]. Samples were then loaded into IrysChips and run on the Irys imaging instrument (BioNano Genomics). A 98-fold coverage (256 Gb) optical map of the sample was produced in two instrument runs with labeled single molecules above 100 kb in size. The IrysView (BioNano Genomics) software package was used to produce single-molecule maps and *de novo* assemble maps into a genome map (**Table 1**).

Scaffolding was also performed using Hi-C-based proximity-guided assembly (PGA). Hi-C libraries were created from goat whole-blood cells (WBC) as described[58]; in this case the sequenced animal was used, as samples were taken before its death. Briefly, cells were fixed with formaldehyde and lysed, and the cross-linked DNA digested with HindIII. Sticky ends were biotinylated and proximity ligated to form chimeric junctions that were enriched for and then physically sheared to a size of 300–500 bp. Chimeric fragments representing the original cross-linked long-distance physical interactions were then processed into paired-end sequencing libraries, and 115 million 100-bp paired-end Illumina reads were produced. The paired-end reads were uniquely mapped onto the draft assembly contigs, which were grouped into 31 chromosome clusters and scaffolded using Lachesis software[15] with tuned parameters (**Supplementary Note**).

Conflict resolution. Our tiered approach to scaffolding provides several opportunities for resolving misassemblies and contig orientation mistakes made by prior steps (for more detail, see **Supplementary Note**). In order to resolve all conflicts from our final assembly, we used a consensus approach that used evidence from five different sources of information: (i) our long-read-based contig sequence, (ii) Irys optical maps, (iii) Hi-C scaffolding orientation quality scores, (iv) San Clemente goat Illumina HiSeq read alignments to the contigs, and (v) a previously generated RH map[34] (**Fig. 1b**). We found that 40 contigs did not align with the Irys optical map, and there were 102 Irys conflicts that needed resolution. A large proportion of the conflicts were identified as forks in the minimum tiling path of contigs superimposed on Irys maps (for example, **Fig. 1b**), but we found that 70 of these conflicts were due to ambiguous contig alignments on two or more Irys maps. Assembly forks are conflict regions in the assembly that arise when ambiguity of sequence makes it equally likely that a contig or scaffold's sequence should continue in two (or more) distinct paths. These ambiguous alignments were due to the presence of segmental duplications or divergent, alternative haplotypes on multiple scaffolds and were discarded. Of the original 102 conflicts, only 36 conflicts had drops in Illumina sequence read depth characteristic of a misassembly, and these were later confirmed by the RH map to be chimeric. The PacBio + PGA assembly (before Irys scaffolding) had 131 scaffolds with orientation conflicts compared to the RH map. The PacBio + Irys + PGA data set had 21 orientation conflicts (consisting of 83 scaffolds) with our RH map. After reordering conflict scaffolds using the RH map information, approximately 84.3% of these orientation conflicts (70/83) were filled by PBJelly, confirming that the RH map orientations for these scaffolds were correct and the PGA orientations were errors. We were unable to find any other data set, apart from the RH map, that accurately predicted which PGA scaffolds contained orientation errors to a high degree of specificity. Since the *C. hircus* X chromosome is acrocentric, our two X chromosome scaffolds do not represent distinct arms of the goat X chromosome and were probably split owing to the requested number of clusters in the proximity-guided assembly algorithm. Still, our recommendation is to use the haploid chromosome count as input to Hi-C scaffolding to avoid false positive scaffold merging. We recommend the use of a suitable genetic or physical map resource, larger input scaffolds into the PGA algorithm, or more frequent cutting restriction enzymes in the generation of Hi-C libraries to avoid or resolve these few remaining errors.

Assembly polishing and contaminant identification. After scaffolding and conflict resolution we ran PBJelly from PBSuite v15.8.24 (ref. 35) with all raw PacBio sequences to close additional gaps. PBJelly closed 681 of 1,439 gaps of at least 3 bp in length. A final round of Quiver[28] was run to correct sequence in filled gaps. It removed 846 contigs with no sequence support, leaving 649 gaps larger than 3 bp. Finally, as P5-C3 chemistry has more errors than P4-C2 or P6-C4 (see Quiver FAQ), we generated 23× coverage of the San Clemente goat individual using 250-bp insert Illumina HiSeq libraries, as mentioned previously, for post-processing error correction and conflict resolution. We aligned reads using BWA[59] (v0.7.10-r789) and SAMtools[60] (v1.2). Using PILON[36], we closed 1 gap and identified and corrected 653,246 homozygous insertions (885,794 bp), 87,818 deletions (127,024 bp), and 34,438 (34,438 bp) substitutions within the assembly that were not present in the Illumina data. This matches the expected error distribution of PacBio data, which has ~5-fold more insertions than deletions[61]. Closer investigation of these data revealed that the majority of insertion events (52.01%) were insertions within a homopolymer run, a known bias of the PacBio chemistry (see URLs). PILON also identified 1,082,330 bases with equal-probability heterozygous substitutions, indicating potential variant sites within the genome.

The final assembly was screened for viral and bacterial contamination using Kraken v0.10.5 (ref. 62) with a database including viral, archaeal, bacterial, protozoa, fungi, and human. A total of 183 unplaced contigs and 1 scaffold

were flagged as contaminant and removed. An additional two unplaced contigs were flagged as vector by NCBI and removed.

Assembly annotation. We employed EVidence Modeler (EVM)[63] to consolidate RNA-seq, cDNA, and protein alignments with *ab initio* gene predictions and the CHIR_1.0 annotation into a final gene set. RNA-seq data included six tissues (hippocampus, hypothalamus, pituitary, pineal, testis, and thyroid) extracted from the domesticated San Clemente goat reference animal and 13 tissues pulled from NCBI Sequence Read Archive (**Supplementary Table 11**). Reads were cleaned with Trimmomatic[64] and aligned to the genome with Tophat2 (ref. 65). Alignments were then assembled independently with StringTie[66] and Cufflinks[67] and *de novo* assembled with Trinity[68]. RNA-seq assemblies were combined and further refined using PASA[63]. Protein and cDNA alignments using exonerate and tblastn with Ensembl data sets of *O. aries*, *B. taurus*, *Equus caballus*, *Sus scrofa*, and *Homo sapiens* as well as NCBI annotation of *C. hircus* and *ab initio* predictions by Braker1 ref. 69 were computed. The CHIR_1.0 annotation coordinates were translated into our coordinate system with the UCSC liftOver tool. All lines of evidence were then fed into EVM using intuitive weighting (RNA-seq > cDNA/protein > *ab initio* gene predictions). Finally, EVM models were updated with PASA.

Gap resolution and repeat analysis. Sequence gap locations were identified from the CHIR_1.0, CHIR_2.0, and ARS1 assembly. In order to identify identical gap regions on different assemblies, we used a simple alignment heuristic (**Supplementary Note**). Briefly, we extracted 500-bp fragments upstream and downstream of each gap region using BEDTOOLS[70] in CHIR_1.0 or CHIR_2.0 and then aligned both fragments to the assembly of comparison (for example, ARS1) using BWA MEM[59]. If (i) both fragments aligned successfully within 10 kb on the same scaffold or chromosome (which was a length greater than 99.6% of all CHIR_1.0 and CHIR_2.0 gaps), (ii) the filled sequence did not map back to a repetitive section on the originating assembly, and (iii) the intervening sequence did not contain ambiguous (N) bases, the gap was considered closed. If fragments aligned to two separate scaffolds or chromosomes, then the region was considered a trans-scaffold break. In cases where one or both fragments surrounding a gap did not align, or if there were two or more ambiguous bases between aligned fragments, the gap was considered open. Gaps were confirmed by two methods. The first method confirmed gaps by checking Illumina WGS read alignments from the sequenced animal to the gap region using SAMtools depth version 1.3 (ref. 60) with read alignment filters as follows: -a -q 30 -Q 40. If one or more bases in the filled region had a read depth <5, the gap was considered unresolved. The second method focused on CHIR_1.0 gaps that were filled by both CHIR_2.0 and ARS1. Briefly, the gap closure region was isolated from CHIR_2.0 and mapped to ARS1 using BWA-MEM v 0.7.12 (ref. 59) with default parameters. Alignments with >14 map quality score (<0.04% likelihood that the alignment is misplaced) to the complementary region in ARS1 indicated a consensus gap closure. Repeats were identified using the RepBase library (release 2015-08-07) with RepeatMasker on the ARS1, CHIR_2.0, UMD3.1 (cattle)[41] and Oarv3.1 (sheep; see URLs) reference assemblies. The "quick" (-q) and "species" (for example, -species goat, -species sheep, -species cow) options were the only deviations from the default. Repeats were filtered by custom scripts if they were <75% of the expected repeat length or were below 60% identity of sequence. Gap comparison images between assemblies were created using NUCmer[71].

Centromeric and telomeric repeat analysis. To identify telomeric sequence we used the 6-mer vertebrate sequence (TTAGGG) and looked for all exact matches in the assembly. We also ran DUST[72] with a window size of 64 and threshold of 20. Windows with at least 10 consecutive identical 6-mer matches (forward or reverse strand) intersecting with low-complexity regions of at least 1,500 bp were flagged as potential telomeric sites and those with >5 kb total length reported. To identify putative centromeric features in our assembly, we used centromeric repetitive sequence for goat from a previously published study[48]. Subsequent alignments of that sequence were used to flag collapsed centromeric sequence in our assembly, identifying three unplaced contigs that contained large portions of the repeat. The contigs were mapped to the assembly, and regions at least 2 kb in length reported as centromeric sites. In all but four cases the telomeric and centromeric sequences were within 100 kb of the contig end (**Supplementary Table 12**). In the cluster corresponding to chromosome 1, the centromeric sequence was at position 40 Mb, confirming a misassembly identified by the RH map. In chromosomes 12 and 13 (clusters 13 and 14, respectively) the centromere was <3 Mb from the end, indicating potential assemblies of the short chromosome arms, though this has not yet been experimentally confirmed.

Fosmid end sequencing and analysis. Sheared genomic DNA was end repaired, and fragments were separated by field-inversion agarose gel electrophoresis. Fragments ranging 38–48 kb were electro-eluted and concentrated using a Microcon-30 centrifugal concentrator. The libraries were created by cloning the DNA into the pNGS FOS vector (Lucigen) with propagation in an *Escherichia coli* DH10B host. End sequence libraries were prepared using a NxSeq 40 kb Mate-Pair cloning kit (Lucigen) and sequenced on a MiSeq (Illumina) using two restriction enzymes (BfaI and RsaI) to generate fosmid end libraries. Approximately 5.2 million and 5.5 million 2 × 250-bp reads were generated from the BfaI and RsaI libraries. Accounting for the expected insert size of the fosmids, the physical coverage of the clones was 40-fold for each library (80-fold total). Reads were screened for vector and bacterial host sequence. Reads were aligned to each reference assembly using BWA MEM[59] with default parameters. Lumpy-SV[43] was used to identify structural variations in the alignment data (**Supplementary Note**).

Statistical analysis. R/Bioconductor was used for all statistical analyses. Spearman's rank order correlation was conducted using the cor.test function in the base R set of utilities, with a two.sided alternative hypothesis. $P < 0.05$ was considered statistically significant.

Code availability. All software versions, links and command line arguments are provided in the **Supplementary Note**. Custom scripts and programs are currently hosted in a GitHub repository at the following link: https://github.com/njdbickhart/GoatAssemblyScripts.

Data availability. The Black Yunan Illumina data were downloaded from Sequence Read Archive (SRA051557). The CHIR_1.0 assembly was downloaded from NCBI (GCA_000317765.1); the CHIR_2.0 assembly was downloaded from NCBI (GCA_000317765.2). The PacBio reads, RNA-seq reads, fosmid end sequences, Illumina WGS reads, and Hi-C library reads that were generated for this study have been deposited in GenBank under accession codes PRJNA290100 and PRJNA340281. Optical map data generated for this study have been deposited in GenBank and are accessible at https://submit.ncbi.nlm.nih.gov/ft/byid/myXc0uq8/goat-merge.cmap and https://submit.ncbi.nlm.nih.gov/ft/byid/ueeq9b8k/rawmolecules.bnx. Intermediary assembly FASTA files, accession numbers, and other miscellaneous information can be found at https://gembox.cbcb.umd.edu/goat/index.html or are available from the corresponding authors upon request.

56. Barrière, A. *et al.* Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. *Genome Res.* **19**, 470–480 (2009).
57. Sayre, B.L. *et al.* Goat breeding in the tropics: development and application of genomic tools in a USAID Feed the Future program. Presented at the 50th Annual Meeting of the Brazilian Society of Animal Science (2013).
58. Burton, J.N., Liachko, I., Dunham, M.J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* **4**, 1339–1346 (2014).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Ross, M.G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
62. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
63. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
64. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

doi:10.1038/ng.3802

65. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
66. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA–seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
67. Trapnell, C. *et al.* Transcript assembly and quantification by RNA–Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
68. Grabherr, M.G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA–Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
69. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA–Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
70. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
71. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
72. Morgulis, A., Gertz, E.M., Schäffer, A.A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).