nature biotechnology

OPEN

Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*)

Yang Dong[1,2,12], Min Xie[3,12], Yu Jiang[1,5,12], Nianqing Xiao[10,12], Xiaoyong Du[4,12], Wenguang Zhang[1,6,12], Gwenola Tosser-Klopp[7], Jinhuan Wang[1], Shuang Yang[3], Jie Liang[3], Wenbin Chen[3], Jing Chen[3], Peng Zeng[3], Yong Hou[3], Chao Bian[3], Shengkai Pan[3], Yuxiang Li[3], Xin Liu[3], Wenliang Wang[3], Bertrand Servin[7], Brian Sayre[11], Bin Zhu[10], Deacon Sweeney[10], Rich Moore[10], Wenhui Nie[1], Yongyi Shen[1,2], Ruoping Zhao[1], Guojie Zhang[3], Jinquan Li[6], Thomas Faraut[7], James Womack[9], Yaping Zhang[1], James Kijas[5], Noelle Cockett[8], Xun Xu[1–3], Shuhong Zhao[4], Jun Wang[3] & Wen Wang[1]

We report the ~2.66-Gb genome sequence of a female Yunnan black goat. The sequence was obtained by combining short-read sequencing data and optical mapping data from a high-throughput whole-genome mapping instrument. The whole-genome mapping data facilitated the assembly of super-scaffolds >5× longer by the N50 metric than scaffolds augmented by fosmid end sequencing (scaffold N50 = 3.06 Mb, super-scaffold N50 = 16.3 Mb). Super-scaffolds are anchored on chromosomes based on conserved synteny with cattle, and the assembly is well supported by two radiation hybrid maps of chromosome 1. We annotate 22,175 protein-coding genes, most of which were recovered in the RNA-seq data of ten tissues. Comparative transcriptomic analysis of the primary and secondary follicles of a cashmere goat reveal 51 genes that are differentially expressed between the two types of hair follicles. This study, whose results will facilitate goat genomics, shows that whole-genome mapping technology can be used for the *de novo* assembly of large genomes.

The domestic goat (*Capra hircus*) is widely reared throughout the world, especially in China, India and other developing countries[1]. Goats serve as an important source of meat, milk, fiber and pelts, and have also fulfilled agricultural, economic, cultural and even religious roles since very early times in human civilization[2]. Evidence indicates that the goat might have been domesticated from two wild Capris (*Capra aegagrus* and *Capra falconeri*) ~10,000 years ago within the Fertile Crescent, and then spread quickly following patterns of human migration and trade[3]. Today, there are >1,000 goat breeds, and >830 million goats are kept around the world according to a report by the UN Food and Agriculture Organization (http://www.fao.org/corp/statistics/en/). In addition to their value as domestic animals, goats are now used as animal models for biomedical research, to investigate the genetic basis of complex traits and in the transgene production of peptide medicines[4,5]. Despite the agricultural and biological importance of goats, breeding and genetics studies have been hindered by the lack of a reference genome sequence. In this work, we combined Illumina next-generation sequencing technology and whole-genome mapping of large DNA molecules to obtain a genome sequence for the domestic goat. We then annotated the genome, and identified rapidly evolving genes. Furthermore, based on an annotated set of goat genes, we generated and compared transcriptomic data from secondary hair follicles (which produce the cashmere fiber) with data from primary hair follicles of the Inner Mongolia cashmere goat, shedding light on the genetic basis of the formation of cashmere fibers.

Whole-genome mapping is an improved high-throughput optical mapping technology. Optical mapping has been used to compare the structures of bacterial genomes[6–8], complete bacterial genome assembly[9,10], assist in bacterial artificial chromosome (BAC) assembly[11] and correct genome assembly errors[12]. Two plant genomes[13,14] have recently been sequenced using BACs assembled through such optical mapping. However, the traditional process for generating optical mapping data involves mostly manual steps, and as a result the primary applications of optical mapping have been in the assembly of bacterial genomes, integration of BAC end sequencing and BAC clone assembly, and iterative correction of assemblies. Although traditional optical mapping methods have been applied successfully, they are complex and have low throughput, primarily because the required DNA extension, image capture and data analysis steps are inefficient. As a result, it has not been possible to generate and handle the massive quantities of optical mapping data that are required for the assembly of a large and complex genome.

To obtain a whole-genome restriction map of the goat, we used an automated, high-throughput whole-genome mapping instrument

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]BGI-Shenzhen, Shenzhen, China. [4]Huazhong Agricultural University, Wuhan, China. [5]CSIRO Livestock Industries, St. Lucia, Australia. [6]Inner Mongolia Agricultural University, Hohhot, China. [7]INRA Laboratoire de Génétique Cellulaire, Castanet-Tolosan, France. [8]Utah State University, Logan, Utah, USA. [9]Texas A&M University, College Station, Texas, USA. [10]OpGen, Inc., Gaithersburg, Maryland, USA. [11]Virginia State University, Petersburg, Virginia, USA. [12]These authors contributed equally to this work. Correspondence should be addressed to W.W. (wwang@mail.kiz.ac.cn), J.W. (wangj@genomics.cn), S.Z. (shzhao@mail.hzau.edu.cn) or X.X. (xuxun@genomics.cn).

Received 10 September; accepted 6 December; published online 23 December 2012; doi:10.1038/nbt.2478

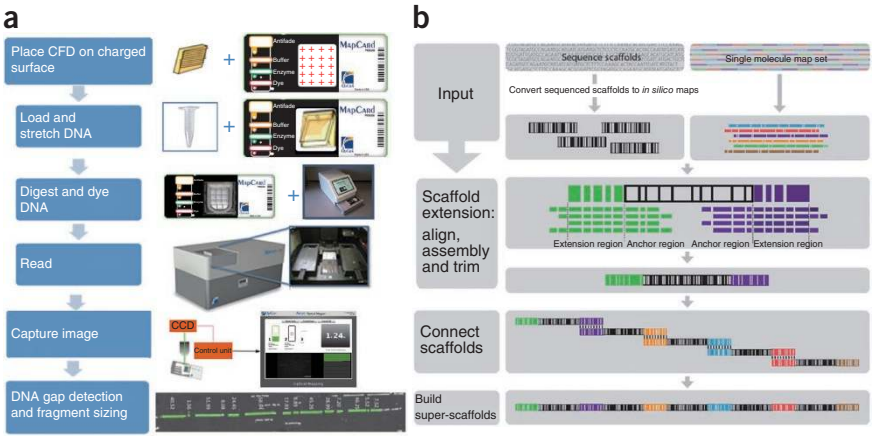

Figure 1 Whole-genome mapping. (**a**) Samples are loaded onto a chip-like, high-density, channel-forming device (CFD). Buffer fluid flowing through the channels stretches high molecular weight DNA onto a positively charged glass surface, which maintains the orientation and integrity of the DNA during subsequent steps. The immobilized single molecules of DNA are digested with a restriction enzyme for 10 min at 37 °C, stained with the dye JOJO-1 and imaged. The images are analyzed channel by channel to filter out nonlinear distorted fragments and small molecules, identify gaps between fragments and measure the size of retained high-quality fragments (colored green) to produce single-molecule restriction maps. (**b**) Scaffolds derived from *de novo* assembly of next-generation sequencing data are converted into restriction maps by *in silico* restriction enzyme digestion. Then, the distance between restriction enzyme sites in the sequencing-derived scaffolds are matched to the lengths of the optical fragments in the single-molecule restriction maps. Matches allow the scaffolds to be extended and linked into super-scaffolds.

and recently developed data processing software. The instrument uses a chip-like channel formation device (CFD) to stretch and immobilize single DNA molecules onto a positively charged glass surface within a disposable cartridge (**Fig. 1a** and **Supplementary Methods**). This, combined with automated imaging and data analysis, addresses many of the inefficiencies that have limited the application of optical mapping to large genomes. The instrument automatically produced 100,000 single-molecule restriction maps in 3 h, providing 12× physical coverage of the goat genome. We then used a hybrid assembly approach to generate super-long scaffolds (super-scaffolds) by combining experimentally measured single-molecule maps with *in silico* restriction maps computed from scaffolds assembled from Illumina sequencing data (**Fig. 1b** and **Supplementary Methods**). The long super-scaffolds facilitated the anchoring of scaffolds onto chromosomes.

RESULTS

Short-read *de novo* sequencing and assembly

We sequenced genomic DNA from a 3-year-old female Yunnan black goat. High-quality DNA extracted from liver tissue was used to construct 14 paired-end sequencing libraries with insert sizes of ~180 bp, 350 bp, 800 bp, 2 kb, 5 kb, 10 kb or 20 kb (**Supplementary Table 1**). Using the Illumina sequencing platform, we generated 191.5 Gb of high-quality reads (65.6-fold coverage of the estimated genome size), with read lengths ranging from 45 to 101 bp (**Supplementary Fig. 1**). These sequences were assembled *de novo* using SOAPdenovo (version 1.03) software[15], resulting in 542,145 contigs and 285,383 scaffolds longer than 100 bp. The contig N50 size was 18.7 kb, which represents the size above which half of the total length of the sequence set can be found. The scaffold N50 size was 2.21 Mb (**Table 1**). To extend the length of scaffolds, we sequenced ends of a fosmid library having an average insert size of ~40 kb constructed from DNA of the same goat (**Supplementary Methods** and **Supplementary Fig. 2**). A total of 2,041,189 paired unique sequences were generated from the fosmid ends, of which 140,296 pairs were matched to different scaffolds and were thus usable for joining scaffolds. This process increased the scaffold N50 size to 3.06 Mb (**Table 1**) and yielded an ~2.66-Gb assembly containing ~140 million *N*s (5.26%) to fill in gaps. The assembled genome is ~91% of the estimated ~2.92-Gb size of the goat genome, based on predictions using the 17-mer method (**Supplementary Fig. 3**). To validate the quality of this assembly, we mapped onto it the raw reads generated from the small insertion libraries, which had been used for contig assembly and gap filling. Over 89% of the raw paired-end reads could be mapped to the assembled goat genome, of which 95% had the correct orientation and correct distance between the ends, indicating that the assembly is largely correct at the local level (**Supplementary Table 2**).

Super-scaffold construction

Information about large linkage groups, such as chromosomes, is important for linkage analysis in animal breeding. Although assembling next-generation sequencing data into a draft genome comprising scaffolds is relatively straightforward, constructing a physical map of the structure of the chromosomes is still difficult and costly. Because a genetic or physical map is not yet available for the goat, we used whole-genome mapping technology to generate a restriction map of the goat genome and then assembled scaffolds into super-scaffolds that were on the order of the length of full chromosomes.

Table 1 Genome assembly statistics

	Contig		Scaffold		Scaffold with fosmid sequences		Super-scaffold	
	bp	Number	bp	Number	bp	Number	bp	Number
N90	4,410	141,869	440,999	1,348	582,523	976	3,825,368	167
N80	7,994	100,335	846,998	922	1,175,001	664	6,481,573	119
N70	11,323	73,948	1,253,003	664	1,739,998	481	9,700,927	88
N60	14,862	54,526	1,694,371	482	2,447,724	352	13,235,657	66
N50[a]	18,720	39,408	2,212,139	344	3,057,189	254	16,328,867	49
Total[b]	2,522,851,955	542,145	2,662,658,003	285,383	2,662,728,047	284,683	2,446,439,202[c]	315

[a]N50 refers to a length-weighted median such that 50% of the genome is contained in contigs or scaffolds of the indicated size or greater. [b]The number of contig or scaffold sequences with length >100 bp. [c]Total length of the super-scaffolds only, excluding small scaffolds that were not assembled into super-scaffolds.

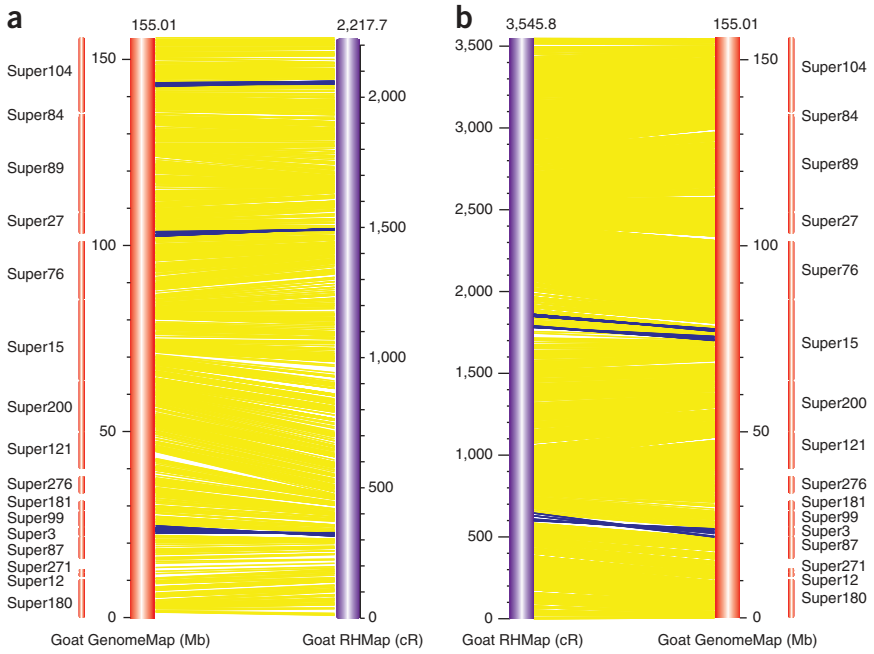

Figure 2 Colinearity between super-scaffolds and the radiation hybrid maps (RHMap) of goat chromosome 1. (**a**,**b**) Maps were generated using BovineSNP50 BeadChip (**a**) and OvineSNP50 BeadChip (**b**). Goat GenomeMap is the assembled pseudo-chromosome 1 generated by anchoring super-scaffolds and scaffolds (regions between super-scaffolds) using the published bovine genome (UMD_3.1 and Batu_4.0). Super-scaffolds anchored on the chromosome are indicated beside the GenomeMap. Only a few rearrangements (blue lines between RHMap and GenomeMap) exist within super-scaffolds.

To obtain single-molecule restriction maps, we used large DNA molecules from a fibroblast cell line established from skin from the ear of the sequenced female Yunnan black goat (**Supplementary Fig. 4**). A total of 3,447,997 single-molecule restriction maps longer than 250 kb each, with an average size of 360 kb, were generated using the SpeI restriction enzyme. The total size of the restriction map data was ~1,241 Gb. A hybrid assembly algorithm, which compares the experimentally determined restriction maps with the *in silico* restriction maps computed from scaffolds assembled from short-read data, was used to identify adjacent scaffolds and determine their relative location and orientation (**Supplementary Methods**). This process joined 2,090 scaffolds, which had an average length of >1.2 Mb, into 315 super-scaffolds. The final assembly had an N50 of 16.3 Mb and covered 92% of assembled scaffolds. The remaining 8% of scaffolds were too small (average length of 713 bp) to be used for whole-genome mapping. The largest super-scaffold was 56.4 Mb (**Table 1**).

To assess the quality of super-scaffolds, we used them to map goat expressed sequence tag (EST) sequences from the NCBI database (http://www.ncbi.nlm.nih.gov/nucest) and assembled *de novo* goat transcriptomes that we obtained from ten tissues (~56 Mb in total). Among the 38,006 ESTs that were >300 bp (average length 1,006.5 bp), 99.2% had hits covering ≥96.3% of their length as revealed with BLAT[16] (version 34, identity >95%) (**Supplementary Methods** and **Supplementary Table 3**).

We also used the core eukaryotic genes mapping approach (CEGMA) pipeline to evaluate the goat assembly[17]. With it, we mapped 97.58% of the core eukaryotic genes (http://korflab.ucdavis.edu/Datasets/cegma/) from six model organisms (*Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Saccharomyces pombe*) to the goat super-scaffolds with coverage >70% (**Supplementary Table 4**). This mapping rate is higher than that obtained for the cattle genome[18], which thus supports the completeness and high quality of the goat super-scaffold assembly.

Anchoring super-scaffolds to the chromosomes

The domestic goat has 29 pairs of autosomes and one pair of sex chromosomes ($2n = 60$)[19]. Cytogenetic comparisons indicate a high level of colinearity between goat and cattle chromosomes, and all 30 goat chromosomes have been ordered according to the International System for Chromosome Nomenclature for Bovids[20]. Based on chromosomal colinearity, we used the two cattle genome assemblies (UMD_3.1 and Btau_4.0) to anchor super-scaffolds to goat chromosomes. Specifically, 302 of the 315 super-scaffolds and 140 other scaffolds that were not included in super-scaffolds were assembled into 30 pseudo-chromosomes for the goat. In total, we anchored 2.52 Gb to the 30 pseudo-chromosomes, and assigned 138 Mb of unordered or unoriented small scaffolds or super-scaffolds to an artificial chromosome designated U. This assembly, which we refer to as CHIR_1.0, is publicly available through a genome browser interface and database (http://goat.kiz.ac.cn/GGD/).

To assess the reliability of the chromosome anchoring, we examined 28 goat genes that have been assigned to a specific chromosomal location[21] (**Supplementary Table 5**). The chromosome assignments of all 28 genes were consistent with our results. As another test of the quality of our pseudo-chromosome assembly, we compared pseudo-chromosome 1 with two radiation hybrid maps of goat chromosome 1 that we generated for a male Boer goat from either 1,222 single-nucleotide polymorphism (SNP) markers on the Illumina BovineSNP50 BeadChip (**Fig. 2a** and **Supplementary Methods**) or 1,567 SNP makers on the Illumina OvineSNP50 BeadChip (**Fig. 2b** and **Supplementary Methods**). We found few rearrangements within super-scaffolds, which were assembled without using cattle colinearity information. In addition, we found few rearrangements between the pseudo-chromosome 1 assembly and the two radiation hybrid maps. Taken together, these results suggest that the assemblies of super-scaffolds and chromosomes are accurate.

We extended our comparison between goat and cattle to all chromosomes. All autosomes were in strong colinearity (**Supplementary Fig. 5a**). Because most goat super-scaffolds are long (N50 = 16.3 Mb), if the super-scaffolds were of low quality, we would expect to see many rearrangements between goat and cattle, but this is not the case (for an example of the high colinearity between a goat and a cattle chromosome, see **Supplementary Fig. 5b**).

Notably, we observed large rearrangements between the X chromosomes of goat and cattle (**Supplementary Fig. 6**), even though the X-chromosome linkage group is usually conserved in placental mammals[21]. The same rearrangements were observed on the X chromosomes when comparing them to both cattle genome assemblies (UMD_3.1 and Btau_4.0). Even within a single goat super-scaffold there are large

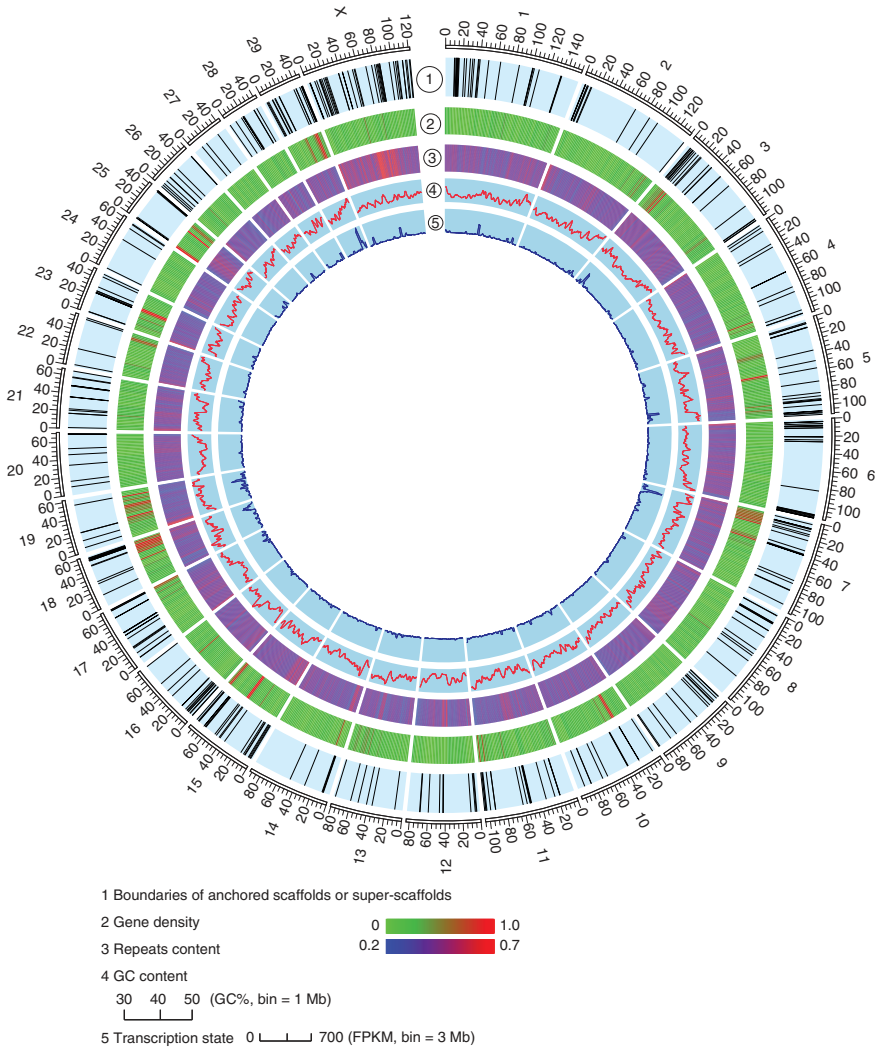

Figure 3 Summary of goat chromosome assemblies. (1) Ideograms of the 30 chromosomes of the goat (in Mb scales). The estimated length of each chromosome is indicated in the outermost ruled circle. Boundaries of anchored super-scaffolds and scaffolds are shown as black lines. (2) Gene density represented as the percentage of the sequence encoding genes for nonoverlapping, 1-Mb windows. (3) Percentage of coverage of repetitive sequences for nonoverlapping, 1-Mb windows. (4) Percentage GC content for nonoverlapping, 1-Mb windows. (5) Transcription state. The transcription level for each gene was estimated by averaging the fragments per kb exon model per million mapped reads (FPKM) from different tissues in nonoverlapping 3-Mb windows.

rearrangements (**Supplementary Fig. 6c**). Because the super-scaffolds were assembled without referring to cattle synteny information, these rearrangements are probably not a result of incorrect assembly based on the optical mapping data, but rather are due to the divergence of the two species. In addition, the sheep reference genome, which was generated by integrating dense physical maps and a large BAC sequence data set (Y. Jiang and International Sheep Genome Consortium, unpublished data), is highly congruent with our goat genome and contains the same rearrangements on the X chromosome. These observations suggest that the large rearrangements between bovine and caprine X chromosomes are real, and support the high quality of our goat genome assembly.

Repetitive sequences and transposable elements

Transposable elements make up a substantial fraction of mammalian genomes and contribute to gene and/or genome evolution[22]. The goat genome has transposable elements similar to those of cattle[22] in that the genome contains large numbers of ruminant-specific repeats, which comprise 42.2% of the goat genome (**Fig. 3** and **Supplementary Table 6**). However, the goat genome has ~80% fewer SINE-BovA repeats (971,273 in goat and 1,839,497 in cattle) and >40% more SINE-tRNA repeats (665,366 in goat and 388,920 in cattle) than does the cattle genome, suggesting that the SINE-BovA repeat expanded primarily in the cattle genome[22], whereas the SINE-tRNA repeat expanded specifically in the goat.

We also analyzed the degree of divergence for each type of transposable element in the goat genome, and found few recently diverged transposable element classes (**Supplementary Fig. 7**). This may be a result of the difficulty of calling repeats with high similarity from short-read sequencing data[23]. However, the general distribution patterns of transposable element classes across chromosomes are similar to that of other assembled mammalian genomes[24,25] (**Supplementary Fig. 8**).

Gene and gene family annotation

We used three gene-prediction methods (homology-based annotation, *ab initio* prediction and RNA-seq-/EST-/cDNA-based annotation) to annotate protein-coding genes. We then merged the results from each method to obtain a consensus gene set of 22,175 protein-coding genes (**Fig. 3**, **Supplementary Tables 7** and **8**), with a mean coding sequence length of 1,385 bp and an average of eight exons per gene. The average lengths of exons and introns were 168 bp and 3,955 bp, respectively (**Supplementary Table 7**). In total, 17,927 annotated protein-coding genes were expressed in at least one of the ten tissues examined by transcriptome sequencing (RNA-seq) (**Fig. 3**, **Supplementary Table 9** and http://goat.kiz.ac.cn/GGD/). Because untranslated regions are difficult to annotate, we used the RNA-seq reads to extend the untranslated regions of 4,740 genes. The gene models derived for goat were highly similar to those of closely related species, supporting the quality of the annotation (**Supplementary Fig. 9**).

We identified 17,129 orthologous gene pairs between goats and cattle, and 16,771 orthologous gene pairs between goats and humans. A phylogenetic tree constructed from 8,325 single-copy orthologs in goats, cattle, horses, dogs, opossums and humans suggests that goats shared a common ancestor with cattle about 23 million years ago (**Fig. 4a**). We further compared orthologous gene pairs between goat and cattle based on ratios of nonsynonymous (*Ka*) and synonymous (*Ks*) substitution rates to identify 44 rapidly evolving genes under positive selection (**Supplementary Table 10**), of which seven are immune-system genes and three are pituitary hormone or related genes. The rapid evolution of immune-system

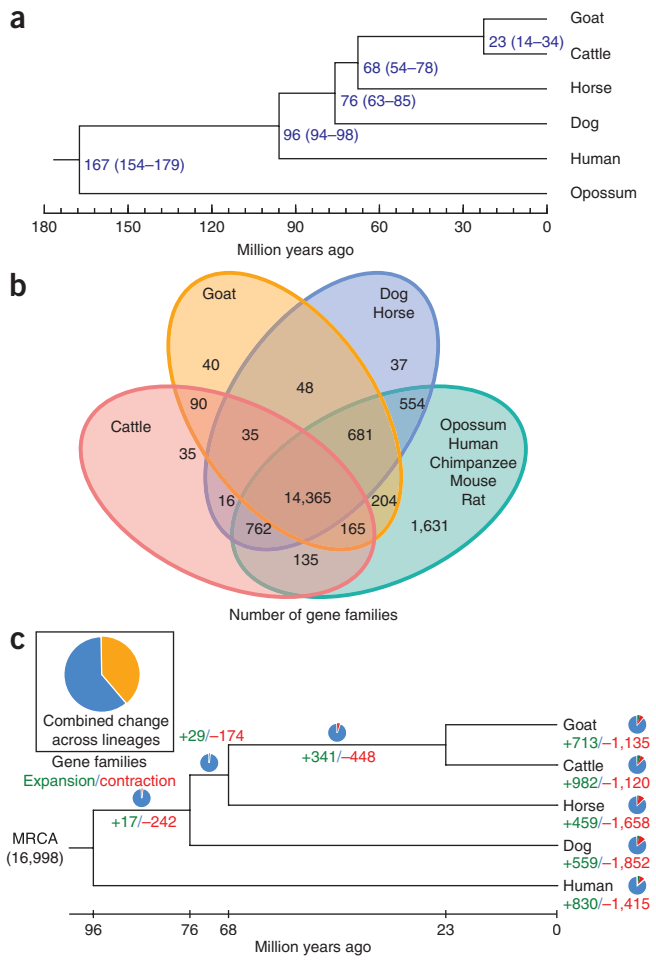

Figure 4 Goat gene family analysis and phylogenetic tree of several mammals. (**a**) Phylogenetic tree constructed with the fourfold degenerate sites of 8,325 single-copy genes. Estimates of divergence time and its interval based on sequence identity are indicated at each node. (**b**) Venn diagram showing the number of unique and shared gene families among nine sequenced mammalian species. (**c**) Dynamic evolution of orthologous gene clusters. The estimated numbers of orthologous groups (16,998) in the most recent common ancestral species (MRCA) are shown at the root node. The numbers of orthologous groups that expanded or contracted in each lineage are shown on the corresponding branch; +, expansion; –, contraction.

genes has also been observed in cattle[24]. Rapid evolution of pituitary hormones may be related to differences between goats and cattle in milk production, development rates of the fetus and/or hair variation, which are traits associated with pituitary hormomes[26,27] (**Supplementary Discussion**).

The assembly contained 262 rRNA, 829 tRNA and 1,010 small nuclear RNA genes (**Supplementary Table 11**). We also identified 487 microRNA (miRNA) genes, of which 157 were located in 44 genomic clusters containing from 2 to 46 miRNA genes (**Supplementary Fig. 10a**). This distribution pattern is similar to that of cattle. However, there are several miRNA gene clusters specific to goats (**Supplementary Fig. 10b**). The largest miRNA gene cluster is located on goat chromosome 21 (**Supplementary Fig. 11**), which is a conserved mammalian miRNA cluster. We used miRNA sequences in other species (human, cattle, dog, chimpanzee, mouse and rat) to identify goat-specific miRNA genes, and found six goat-specific miRNA genes in total (**Supplementary Table 12**), which have typical miRNA structures (**Supplementary Fig. 12**) and many target genes (**Supplementary Fig. 13**).

Based on pair-wise protein sequence similarity, we carried out a gene family clustering analysis on all goat genes compared to genes in cattle, horse, dog, mouse, rat, opossum, chimpanzee and human. The 19,607 goat genes could be clustered into 15,628 gene families (**Fig. 4b** and **Supplementary Table 13**). We identified 40 goat-specific gene families that contain 106 genes; and 43 of these genes were expressed in the ten sequenced tissues (**Supplementary Table 14**). Of the 115 genes found within the 90 ruminant-specific gene families, 68 were expressed (**Supplementary Table 15**). These lineage-specific gene families may have contributed specifically to the evolution of goats or ruminants.

We also analyzed gene-family expansion or contraction events in the goat, cattle, horse, dog and human. In all five genomes, we found a higher frequency of gene-contraction than gene-expansion events (**Fig. 4c**), which has been noted previously[28]. We focused on the most significant expansion or contraction events ($P < 0.01$), and after manually filtering out gene families whose members had different function assignments (**Supplementary Table 16**), we detected three expanded olfactory receptor gene subfamilies but only one contracted subfamily in the goat compared with the cattle, horse, dog and human. It is possible that the olfactory receptor expansion events may contribute to the exceptional foraging ability of goats[29]. We also noticed expansion of the ferritin heavy chain gene (FTH1) family in the goat, with the number of goat FTH1 genes nearly seven times that of the human and two times of the cattle. The expansion of FTH1 in goats may account for its unusual detoxification ability and thus its broad forage diet, as ferritin plays a major role in iron sequestration, detoxification and storage[30].

Two contigs containing the sheep major histocompatibility (MHC) loci (also designated as ovine lymphocyte antigen or OLA) generated with BAC-by-BAC sequencing[31] were used to search against the CHIR_1.0 for goat MHC loci. As expected, the goat MHC loci were located on chromosome 23 in our assembly. Similar to the sheep MHC, the goat MHC contains two regions 2.25 Mb and 360 kb in length, respectively (**Fig. 5a**). Based on a comparison to the sheep MHC, which contains 177 genes, we annotated 160 protein-coding genes (**Supplementary Table 17**) using the same method as for the sheep MHC annotation[31]. Based on the annotation, we also analyzed conserved genes of MHC loci in sheep, goat and human. Even though there are some inversions, which are common for MHC loci, most of the conserved genes show high colinearity among goat, sheep and human (**Fig. 5b**). These results not only indicate that our assembly of the goat genome is of good quality, but also provide a detailed map for the goat MHC, which will be useful for immunological studies and vaccine development.

Transcriptomes of primary and cashmere hair follicles

Mammalian hair is a highly keratinized tissue produced by hair follicles within the skin. There are two kinds of hair follicles: the primary hair follicle produces the coarse coat hair in all mammals, and the secondary hair follicle can produces the cashmere or 'fine hair' in certain mammals, including goats and antelopes[32] (**Supplementary Fig. 14**). Characterized by its fine and soft features, cashmere fiber has been obtained mainly from the cashmere goat. Despite a 2,500-year history and enormous raw cashmere production, estimated at some 10,000 tons per year in China, the world's largest producer of cashmere[33], little is known about the molecular mechanisms of cashmere formation and development.

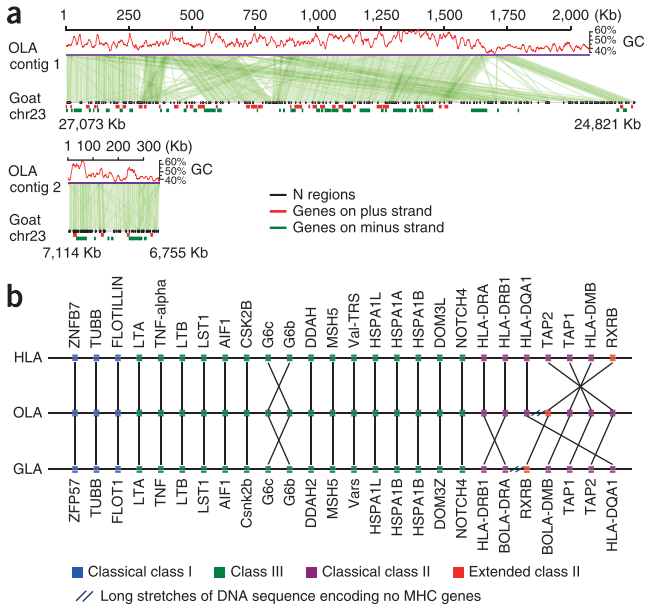

Figure 5 Comparative analysis of the goat MHC. (**a**) A map of the goat MHC and colinearity of the goat MHC with the sheep MHC (OLA). Green lines show syntenic relationship between the goat MHC and OLA. Only genes within the goat MHC are marked. Also shown, GC content (%, with nonoverlapping, 1 kb windows). (**b**) Genes conserved between goat MHC (GLA), sheep MHC (OLA) and human MHC (HLA) are connected by black lines.

We investigated the genetic basis underlying the development of cashmere fibers by sequencing the transcriptomes of primary and secondary hair follicles and mapping the reads to the goat genome assembly and annotated genes. RNA was extracted from 20–50 secondary or primary hair follicles of an Inner Mongolia cashmere goat, yielding 144–588 ng RNA per sample, and the transcriptomes of three pairs of primary and secondary hair follicle samples (three biological replicates) were directly sequenced without amplification, generating 20.3 Gb of sequence data (**Supplementary Table 18**). The majority (~75%) of the total FPKM (fragments per kilobase of exon per million fragments mapped) values in both hair follicles were from keratin and keratin-associated protein genes (**Supplementary Table 19**). Across all three paired samples, we identified 10,077 genes in the primary hair follicle samples and 7,772 genes in the secondary hair follicle samples with FPKM > 0.1. Of the 2,572 genes in the primary hair follicle samples and 1,947 genes in the secondary hair follicle samples with FPKM > 5, 51 showed a change in expression of at least twofold between all three pairs of secondary and primary hair follicle samples (**Supplementary Table 20**), with 28 downregulated and 23 upregulated in secondary versus primary follicles.

Keratin and keratin-associated proteins are the main structural proteins of hair fibers, determining the quality of fiber. Two types (type I and type II) of keratins are paired to form obligatory heteropolymers[34], whereas the keratin-associated proteins may be responsible for forming the rigid hair shaft and altering the hair structure and diameter[35]. In total, we annotated 49 keratin genes (**Supplementary Table 21**) and 30 keratin-associated protein genes in the goat genome (**Supplementary Table 22**), of which 29 keratin genes and all 30 keratin-associated protein genes were detected with FPKM > 5 in both types of follicles (**Supplementary Table 23**). Notably, two of the 29 keratin genes and 10 of the 30 keratin-associated protein genes were consistently differentially expressed between primary and secondary hair follicles in all three sample sets, and all of these were expressed higher in secondary than in primary follicles (**Supplementary Table 24**), suggesting that the keratin-associated protein genes may be more important in determining the structure of cashmere fibers. The two differentially expressed keratin genes (keratin 40 and 72) were type 1 and type 2, respectively (**Supplementary Fig. 15**). Keratin-associated proteins can be divided into three major groups: high sulfur, ultra-high sulfur and high glycine-tyrosine[36]. The ten differentially expressed keratin-associated proteins were all in the ultra-high sulfur group (**Supplementary Fig. 16**), suggesting that this group of proteins may be important for the formation of cashmere.

Other upregulated genes in secondary hair follicles include fibroblast growth factor 21 (GOAT_ENSP00000222157), which can promote the transition to catagen[37], and casein kinase Iε(goat_GLEAN_10015556), an important regulator of β-catenin in the Wnt pathway, which is one of the most important pathways in hair follicle development[38].

The downregulated genes in secondary hair follicles included two enzymes of amino acid biosynthesis, asparagine synthetase (goat_GLEAN_10019946) and phosphoserine aminotransferase (GOAT_ENSP00000388939), which are key enzymes in asparagine and serine biosynthesis, suggesting these amino acids could be more intensively involved in primary hair growth. Other downregulated genes included Gap junction alpha-1 protein (goat_GLEAN_10013034) and Desmoglein 1 (GOAT_ENSBTAP00000018382), which have been reported to be involved in hair follicle cell communication and hair follicle morphogenesis[38,39], and isopentenyl-diphosphate delta-isomerase 1 (GOAT_ENSBTAP00000005323) and cellular retinoic acid–binding protein 2 (GOAT_ENSBTAP00000007515), which are both related to retinoic acid biosynthesis and can regulate hair growth and the hair life cycle through Wnt signaling[40,41]. Further analyses of these expression data generated additional hypotheses related to genes and pathways that may underlie cashmere fiber production (**Supplementary Discussion**, **Supplementary Fig. 17** and **Supplementary Tables 25–29**).

DISCUSSION

The goat genome is, to our knowledge, the first large genome to be sequenced and assembled *de novo* using whole-genome mapping technology, demonstrating that this approach can be used to obtain a highly contiguous assembly for a large genome without the aid of traditional genetic maps. The long super-scaffolds provide sufficient linkage-group information for gene mapping and marker-assisted breeding, and they are long enough to be anchored onto chromosomes using rough colinearity information of other closely related mammals whose complete genomes are available. We plan to update the goat genome assembly as radiation hybrid maps for all chromosomes become available.

The goat genome sequence will be useful for mapping reads obtained by resequencing more breeds of goats, which will facilitate the identification of SNP markers for genomic marker–assisted breeding. To our knowledge, the goat is the first small ruminant whose genome has been sequenced. The goat genome should be useful for understanding the genomic features that distinguish ruminants from nonruminant species. It will also be useful for improving the utility of the goat as a biomedical model and bioreactor. In addition, the genes we identified that are related to cashmere fiber production could be used as markers for breeding better cashmere goats, or they may be potential targets for genetic or nongenetic manipulation.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. The goat whole-genome shotgun project, DDBJ/EMBL/GenBank: AJPT00000000. The version described here is AJPT00000000.1. DNA sequencing short reads, SRA: SRA051557. RNA sequencing short reads, GEO: GSE37456. Because there is no data bank for storing optical mapping data yet, the whole-genome mapping data of goat can be obtained from http://goat.kiz.ac.cn/GGD/. The radiation hybrid map data of chromosome 1 are also available from http://goat.kiz.ac.cn/GGD/.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the key project of CAS (KSZD-EW-Z-005), Chinese 973 programs (2007CB815700), the National Natural Science Foundation of China (30960246 and 31260538), the Shenzhen Municipal Government and the Yantian District local government of Shenzhen, and a CAS-Max Planck Society Fellowship and the 100 talent program of CAS to W.W. and Jun W.

AUTHOR CONTRIBUTIONS

Y.D., Wen W., Jun W., X.X., W.Z., N.C., James W., G.T.-K., T.F., J.K., Bertrand S., Brian S. and Y.Z. planned and coordinated the project, and wrote the manuscript; Y.D., W.Z., Jinhua W., W.N., R.Z. and J.L. prepared samples; Y.J., X.D, M.X., J.L., W.C. and Y.S. mapped and assembled the short reads; N.X., B.Z., D.S. and R.M. conducted the whole-genome mapping and related computational analysis; Y.D., X.X., Wenliang W., S.Y., J.L., J.C., P.Z., G.Z., Y.L. and M.X. sequenced and processed the RAW data; Y.J., M.X., Y.D., G.Z., X.L., Brian S. and J.K. annotated the genome; X.D., S.Z. and T.F. provided the radiation hybrid map; M.X. and C.B. analyzed gene family; Y.D., Y.J., S.P. and Y.S. conducted positive selection analysis; Y.H., Y.Z., N.C. and Jinhua W. constructed the pseudo-chromosomes.

COMPETING FINANCIAL INTERESTS

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

Published online at http://www.nature.com/doifinder/10.1038/nbt.2478.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Zeder, M.A. & Hesse, B. The initial domestication of goats (Capra hircus) in the Zagros mountains 10,000 years ago. *Science* **287**, 2254–2257 (2000).
2. MacHugh, D.E. & Bradley, D.G. Livestock genetic origins: goats buck the trend. *Proc. Natl. Acad. Sci. USA* **98**, 5382–5384 (2001).
3. Luikart, G. *et al.* Multiple maternal origins and weak phylogeographic structure in domestic goats. *Proc. Natl. Acad. Sci. USA* **98**, 5927–5932 (2001).
4. Ebert, K.M. *et al.* Transgenic production of a variant of human tissue-type plasminogen activator in goat milk: generation of transgenic goats and analysis of expression. *Bio/Technology* **9**, 835–838 (1991).
5. Ko, J.H. *et al.* Production of biologically active human granulocyte colony stimulating factor in the milk of transgenic goat. *Transgenic Res.* **9**, 215–222 (2000).
6. Anantharaman, T., Mishra, B. & Schwartz, D. Genomics via optical mapping. III: Contiging genomic DNA. *International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology* 18–27 (1999).
7. Chen, Q., Savarino, S.J. & Venkatesan, M.M. Subtractive hybridization and optical mapping of the enterotoxigenic *Escherichia coli* H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of *E. coli* K-12. *Microbiology* **152**, 1041–1054 (2006).
8. Lim, A., *et al.* Shotgun optical maps of the whole *Escherichia coli* O157:H7 genome. *Genome Res.* **11**, 1584–1593 (2001).
9. Nagarajan, N., Read, T.D. & Pop, M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **24**, 1229–1235 (2008).
10. Wu, C.W., Schramm, T.M., Zhou, S., Schwartz, D.C. & Talaat, A.M. Optical mapping of the Mycobacterium avium subspecies paratuberculosis genome. *BMC Genomics* **10**, 25 (2009).
11. Wei, F. *et al.* The physical and genetic framework of the maize B73 genome. *PLoS Genet.* **5**, e1000715 (2009).
12. Zhou, S. *et al.* Validation of rice genome sequence by optical mapping. *BMC Genomics* **8**, 278 (2007).
13. Young, N.D. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011).
14. Zhou, S. *et al.* A single molecule scaffold for the maize genome. *PLoS Genet.* **5**, e1000711 (2009).
15. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
16. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
17. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
18. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297 (2009).
19. Wurster, D.H. & Benirschke, K. Chromosome studies in the superfamily Bovoidea. *Chromosoma* **25**, 152–171 (1968).
20. Cribiu, E.P. *et al.* International system for chromosome nomenclature of domestic bovids (ISCNDB 2000). *Cytogenet. Cell Genet.* **92**, 283–299 (2001).
21. Iannuzzi, L., King, W.A. & Di Berardino, D. Chromosome evolution in domestic bovids as revealed by chromosome banding and FISH-mapping techniques. *Cytogenet. Genome Res.* **126**, 49–62 (2009).
22. Adelson, D.L., Raison, J.M. & Edgar, R.C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc. Natl. Acad. Sci. USA* **106**, 12855–12860 (2009).
23. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
24. Elsik, C.G. *et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–528 (2009).
25. Wade, C.M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–867 (2009).
26. Zhang, D. *et al.* Cow placenta extract promotes murine hair growth through enhancing the insulin-like growth factor-1. *Indian J. Dermatol.* **56**, 14–18 (2011).
27. Handwerger, S. & Freemark, M. The roles of placental growth hormone and placental lactogen in the regulation of human fetal growth and development. *J. Pediatr. Endocrinol. Metab.* **13**, 343–356 (2000).
28. Olson, M.V. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* **64**, 18–23 (1999).
29. Gilad, Y., Przeworski, M. & Lancet, D. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol.* **2**, e5 (2004).
30. Theil, E.C. Ferritin: structure, gene regulation, and cellular function in animals, plants, and microorganisms. *Annu. Rev. Biochem.* **56**, 289–315 (1987).
31. Gao, J. *et al.* A complete DNA sequence map of the ovine major histocompatibility complex. *BMC Genomics* **11**, 466 (2010).
32. Ibraheem, M., Galbraith, H., Scaife, J. & Ewen, S. Growth of secondary hair follicles of the Cashmere goat *in vitro* and their response to prolactin and melatonin. *J. Anat.* **185**, 135–142 (1994).
33. White, C. *The Cashmere Shawl; An Eastern Fiction Volume 1* (RareBooksClub.com, 1840).
34. Sun, T.T. *et al.* Keratin classes: molecular markers for different types of epithelial differentiation. *J. Invest. Dermatol.* **81**, 109s–115s (1983).
35. Rogers, M.A., Langbein, L., Praetzel-Wunder, S., Winter, H. & Schweizer, J. Human hair keratin-associated proteins (KAPs). *Int. Rev. Cytol.* **251**, 209–263 (2006).
36. Rogers, M.A. & Schweizer, J. Human KAP genes, only the half of it? Extensive size polymorphisms in hair keratin-associated protein genes. *J. Invest. Dermatol.* **124**, vi–ix (2005).
37. Schlake, T. FGF signals specifically regulate the structure of hair shaft medulla via IGF-binding protein 5. *Development* **132**, 2981–2990 (2005).
38. Sakanaka, C., Leong, P., Xu, L., Harrison, S.D. & Williams, L.T. Casein kinase iepsilon in the wnt pathway: regulation of beta-catenin function. *Proc. Natl. Acad. Sci. USA* **96**, 12548–12552 (1999).
39. Hanakawa, Y., Li, H., Lin, C., Stanley, J.R. & Cotsarelis, G. Desmogleins 1 and 3 in the companion layer anchor mouse anagen hair to the follicle. *J. Invest. Dermatol.* **123**, 817–822 (2004).
40. Lefebvre, P.P., Malgrange, B., Staecker, H., Moonen, G. & Van de Water, T.R. Retinoic acid stimulates regeneration of mammalian auditory hair cells. *Science* **260**, 692–695 (1993).
41. Viallet, J.P. & Dhouailly, D. Retinoic acid and mouse skin morphogenesis. II. Role of epidermal competence in hair glandular metaplasia. *Dev. Biol.* **166**, 277–288 (1994).

ONLINE METHODS

DNA/RNA isolation, library construction and sequencing. Genomic DNA was isolated from liver tissue of a female Yunnan black goat by standard molecular biology techniques. DNAs were sheared to fragments of 180–800 bp, 2 kb, 5 kb, 10 kb and 20 kb to generate the PE libraries (see **Supplementary Methods** for details). All these DNA libraries were sequenced on the Illumina Genome Analyzer II platform.

High-quality DNA extracted from liver tissue of the female Yunnan black goat was used for the fosmid library construction (see **Supplementary Methods** for details). Fosmid end sequencing was done in this order: fragmentation and end repair, size selection and purification, circularization of digested linear DNA, inverse PCR and enrichment of 400–700 bp DNA fragments (see **Supplementary Methods** for details). Illumina sequencing for short insert libraries was then done.

RNA was purified using TRIzol (Invitrogen). RNA sequencing libraries were constructed using the mRNA-Seq Prep Kit (Illumina, USA). We sequenced 200 bp paired-end libraries of RNA-seq using the paired-end sequencing module (90 bp at each end) of the Illumina HiSeq 2000 platform (see **Supplementary Methods** for details).

Scaffold construction. Goat genome scaffolds were constructed using SOAPdenovo software (Release 1.03, http://soap.genomics.org.cn/, parameter "-K 41 -d 1 -M 2 -F," see **Supplementary Methods** for details). The end sequences from the fosmid library were used to extend the scaffolds using the procedure described in **Supplementary Methods.**

Whole-genome mapping (WGM) and construction of super-scaffolds. We used the new WGM technology developed by the Argus System and WGM software package (Genome-Builder) of OpGen to produce huge optical mapping data, process these data and extend the scaffolds fully automatically. The system integrates wet lab chemistry, including digestion and staining, into an automated process using MapCard and MapCard Processor, followed by automatically collection of over 7,000 fluorescence microscope images per MapCard by the Argus Mapper instrument. The efficient version of Genome-Builder for small data sets is embedded in the Argus System, but for large data sets the Genome-Builder needs to be installed in a computer server.

High molecular weight DNA from fibroblast cell line from skin tissue of the female Yunnan black was passed through a channel-forming device (CFD) to direct and stretch the single DNA molecules onto a positively charged glass surface in MapCard, which had separate chambers for all reagents to be preloaded (see **Supplementary Methods** for details). DNA was elongated and immobilized to the surface after it flowed down through the micro channels of CFD. Fixing the DNA to the surface prevented recoiling, ensuring optimal orientation of the DNA for image capture by the CCD camera. The immobilized single molecules of DNA were digested with SpeI for 10 min at 37 °C and subsequently stained with JOJO-1 (Life Technologies) on the MapCard Processor (MCP) (**Supplementary Fig. 4a**). The MCP automates the sequential restriction enzyme digestion and staining steps.

Individual DNA molecules and corresponding restriction fragments were imaged by laser-illuminated fluorescent microscopy using the Argus Mapper (see **Supplementary Methods** for details). The restriction enzyme cut sites were detected as gaps in DNA images, and the size of each restriction fragment between adjacent cut sites was determined (**Supplementary Fig. 4c**). The Mapper analyzes the images channel by channel, filters out nonlinear distorted fragments and small molecules, identifies gaps between fragments and measures size of retained high-quality fragments. For this project, 3,447,997 single-molecule restriction maps (>250 kb) with an average size of 360 kb were generated. The total size of single-molecule restriction map data was about 1,241 Gb.

Super-scaffolding with WGM data. Super-scaffolding with WGM data was performed using Genome-Builder software recently developed at OpGen. This software suite takes a hybrid approach to perform long-range scaffolding of *de novo* sequence assembly. Briefly, it uses single-molecule maps generated in Argus to extend sequence scaffolds, create overlapping regions between adjacent scaffolds and connect the scaffolds based on pair-wise alignments between them. The input sequence scaffolds were based on the *de novo* assembly and were first converted into restriction maps by *in silico* restriction enzyme digestion. The resulting *in silico* maps were used as initial seed maps for an iterative extension process. The details of the algorithm are further described in **Supplementary Figure 4** and **Supplementary Methods**.

The super-scaffolds were evaluated by CEGMA[17] and ESTs downloaded from NCBI (13,849 records) and *de novo* assembled ESTs from RNA-seq reads of ten tissues (99,707 records, see **Supplementary Methods** for details).

Pseudo-chromosomes assembly and evaluation. A set of 108,850 source sequences for SNP probes from the OvineSNP50 BeadChip and BovineSNP50 BeadChip were compared for similarity to the goat super-scaffolds/scaffolds and the cattle genome (UMD 3.1) with BLASTN to locate the super-scaffolds and other scaffolds unmapped in the whole-genome mapping into the 30 pseudo-chromosomes (see **Supplementary Methods** for details). The WGM data were then used to double-check the order of anchored super-scaffolds or scaffolds. Dislocations confirmed by our experimental WGM data were candidates of true alignment differences caused by noncolinearity between the goat and cattle. Files of scaffolds, super-scaffolds and their corresponding .agp instruction files are available at http://goat.kiz.ac.cn/GGD/.

To evaluate the chromosomal assembly of the goat genome, we compared the pseudo-chromosome 1 with two radiation hybrid maps of goat chromosome 1 generated by us for a male Boer goat. We genotyped 1,222 SNP markers on the Illumina BovineSNP50 BeadChip and 1,567 SNP makers on the Illumina OvineSNP50 BeadChip that could be represented in the goat genome across a 5,000 rad goat-hamster radiation hybrid panel that contained 93 cell lines of a Boer goat. We also conducted synteny-based chromosomal comparisons for all chromosomes between goat CHIR_1.0 assembly and both bovine Btau_4.0 and UMD3.1 assemblies (see **Supplementary Methods** for details).

Genome annotations and analyses. Genome annotations include the annotation of repeat elements (transposable elements), protein-coding genes, non-protein-coding genes and gene families. Based on the genome annotations, genome analyses focused on genes under positive selection and gene family evolution. The brief methods were listed here, and details are fully described in **Supplementary Methods**.

Tandem repeats in the genome assembly were identified using Tandem Repeat Finder[42]. Noninterspersed repeats in the genome were detected by using RepeatMasker[43]. Transposable elements in the genome assembly were identified at the DNA and protein levels. At the DNA level, RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) and LTR_FINDER[44] software were used to build *de novo* repeat libraries. RepeatMasker (version 3.2.9) was run against *de novo* library and repbase[45] separately to identify homologous repeats, which were classified into known classes of repeats[46]. At the protein level, RM-BLASTX within RepeatProteinMask in RepeatMasker software package was used against the transposable elements protein database.

To predict protein-coding genes, information was integrated from three different methods, *ab initio* prediction, homology-based annotation and RNA-seq-/EST-/cDNA-based annotation. RNA-seq data were used to extend the gene sequences, especially 5′ UTRs where high GC content usually hinders Illunima sequencing.

InterProScan[47] (version 4.5) was used to screen goat proteins against five databases (Pfam[48], release 24.0; PRINT[49], release 40.0; PROSITE[50], release 20.52; ProDom[51], 2006.1; MART, release 6.0). The KEGG[52,53] (Release 58), Uniprot/SwissProt[54] (Release 2011.6) and UniProt/TrEMBL[55] (Release 2011.6) database were searched for homology-based function assignments (GO assignments).

The tRNAscan-SE[56] (version 1.23) software with default parameters for eukaryote was used for tRNA annotation. rRNA annotation was based on homology information of human rRNA collections using BLASTN (version 2.2.21). The miRNA and small nuclear RNA genes were predicted by INFERNAL[57] software against the Rfam database[58].

The Treefam[59] methodology was used to define a gene family as a group of genes descending from a single gene in the common ancestor of goat, cattle, horse, dog, mouse, rat, opossum, chimpanzee and human (*Ensembl* release 56).

Single-copy genes defined as orthologous genes by Treefam pipelines were chosen for phylogenetic analysis with MrBayes software[60]. The Bayesian Relaxed Molecular Clock (BRMC) approach was used to estimate the species

doi:10.1038/nbt.2478

divergence time using the program MCMCTREE (version 4), which was part of the PAML package[61]. The divergence time of human and dog from TimeTree database (http://www.timetree.org/) was used as the calibration time.

CAFE (computational analysis of gene family evolution, version 2.1)[62] was used to detect gene family expansion and contraction in human, dog, horse, cattle and goat, respectively. *Ka*/*Ks* ratios were calculated for 14,906 orthologous pairs among goat, cattle and human using *KaKs*_Calculator[63] software and positive selection were further tested these gene pairs.

Comparison between primary hair follicle and secondary hair follicle transcriptomes. After filtering low-quality/contaminated/PCR artifacts reads, reads from RNA-seq data of primary hair follicle (PHF) and secondary hair follicle (SHF) paired samples were mapped against the goat assembly using Tophat[64]. FPKM value was calculated for each protein-coding gene by Cuffdiff (http://cufflinks.cbcb.umd.edu). The significance level (*P*-value) of differential expressed genes between two samples was calculated with Cuffdiff using default parameters. FPKM > 5 was used as the stringent cutoff to identify expressed genes, respectively. Differentially expressed genes are those with at least twofold FPKM change between PHF and SHF samples in all three PHF/SHF comparisons.

42. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
43. Bedell, J.A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040–1041 (2000).
44. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
45. Kapitonov, V.V. & Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **9**, 411–412 author reply 414 (2008).
46. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
47. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.* **396**, 59–70 (2007).
48. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
49. Attwood, T. K. *et al.* The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford)* **2012**, bas019 (2012).
50. Bairoch, A. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19** (suppl.), 2241–2245 (1991).
51. Ribeiro, E.O. *et al.* A distributed computation of Interpro Pfam, PROSITE and ProDom for protein annotation. *Genet. Mol. Res.* **4**, 590–598 (2005).
52. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
53. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).
54. The Uniprot Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142–D148 (2010).
55. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
56. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
57. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
58. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
59. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006).
60. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
61. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
62. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
63. Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).
64. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).