

# SCIENTIFIC REPORTS



OPEN

## Genome-wide Target Enrichment-aided Chip Design: a 66 K SNP Chip for Cashmere Goat

Xian Qiao<sup>1</sup>, Rui Su<sup>1,2,3,4,5</sup>, Yang Wang<sup>6,7</sup>, Ruijun Wang<sup>1,2,3,4</sup>, Ting Yang<sup>6,7</sup>, Xiaokai Li<sup>1</sup>, Wei Chen<sup>10,17</sup>, Shiyang He<sup>7</sup>, Yu Jiang<sup>9</sup>, Qiwu Xu<sup>6,8</sup>, Wenting Wan<sup>11</sup>, Yaolei Zhang<sup>6,8</sup>, Wenguang Zhang<sup>1,2</sup>, Jiang Chen<sup>6,7</sup>, Bin Liu<sup>12</sup>, Xin Liu<sup>6,8</sup>, Yixing Fan<sup>1</sup>, Duoyuan Chen<sup>7</sup>, Huaizhi Jiang<sup>13</sup>, Dongming Fang<sup>7</sup>, Zhihong Liu<sup>1,2,3,4</sup>, Xiaowen Wang<sup>7</sup>, Yanjun Zhang<sup>1,2,3,4</sup>, Danqing Mao<sup>7</sup>, Zhiying Wang<sup>1,2,3,4</sup>, Ran Di<sup>14</sup>, Qianjun Zhao<sup>14</sup>, Tao Zhong<sup>15</sup>, Huanming Yang<sup>7,16</sup>, Jian Wang<sup>7,16</sup>, Wen Wang<sup>5</sup>, Yang Dong<sup>7,10,17</sup>, Xiaoli Chen<sup>6,7</sup>, Xun Xu<sup>6,7</sup> & Jinquan Li<sup>1,2,3,4</sup>

Compared with the commercially available single nucleotide polymorphism (SNP) chip based on the Bead Chip technology, the solution hybrid selection (SHS)-based target enrichment SNP chip is not only design-flexible, but also cost-effective for genotype sequencing. In this study, we propose to design an animal SNP chip using the SHS-based target enrichment strategy for the first time. As an update to the international collaboration on goat research, a 66 K SNP chip for cashmere goat was created from the whole-genome sequencing data of 73 individuals. Verification of this 66 K SNP chip with the whole-genome sequencing data of 436 cashmere goats showed that the SNP call rates was between 95.3% and 99.8%. The average sequencing depth for target SNPs were 40X. The capture regions were shown to be 200 bp that flank target SNPs. This chip was further tested in a genome-wide association analysis of cashmere fineness (fiber diameter). Several top hit loci were found marginally associated with signaling pathways involved in hair growth. These results demonstrate that the 66 K SNP chip is a useful tool in the genomic analyses of cashmere goats. The successful chip design shows that the SHS-based target enrichment strategy could be applied to SNP chip design in other species.

The first-generation animal breeding strategies for important quantitative traits relied heavily on keeping meticulous documentation of animal phenotypes and breeding values over several generations<sup>1</sup>. Because this process was expensive and time consuming<sup>2</sup>, breeding scientists tried to find more efficient methods to select desirable genetic traits. Nearly 35 years ago, Geldermann proposed the concept of quantitative trait loci (QTL) in animal

<sup>1</sup>College of Animal Science, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, 010018, China. <sup>2</sup>Key Laboratory of Animal Genetics, Breeding and Reproduction, Inner Mongolia Autonomous Region, Hohhot, 010018, China. <sup>3</sup>Key Laboratory of Mutton Sheep Genetics and Breeding, Ministry of Agriculture, Hohhot, 010018, China. <sup>4</sup>Engineering Research Center for Goat Genetics and Breeding, Inner Mongolia Autonomous Region, Hohhot, 010018, China. <sup>5</sup>State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, 650223, China. <sup>6</sup>China National GeneBank, BGI-Shenzhen, Shenzhen, 518120, China. <sup>7</sup>BGI-Shenzhen, Shenzhen, 518083, China. <sup>8</sup>BGI-Qingdao, Qingdao, 266000, China. <sup>9</sup>College of Animal Science and Technology, Northwest A&F University, Yangling, 712100, China. <sup>10</sup>College of Biological big data, Yunnan Agriculture University, Kunming, 650504, China. <sup>11</sup>Center for Ecological and Environmental Sciences, Key Laboratory for Space Bioscience & Biotechnology, Northwestern Polytechnical University, Xi'an, 710072, China. <sup>12</sup>Institute of Animal Husbandry, Academy of Agriculture and Stockbreeding Sciences, Hohhot, Inner Mongolia, 010030, China. <sup>13</sup>College of Animal Science and Technology, Jilin Agricultural University, Changchun, 130118, China. <sup>14</sup>The Key Laboratory for Farm Animal Genetic Resources and Utilization of Ministry of Agriculture of China, Institute of Animal Science (IAS), Chinese Academy of Agricultural Sciences (CAAS), Beijing, 100193, China. <sup>15</sup>Institute of Animal Genetics and Breeding, Sichuan Agricultural University, Chengdu, 611130, China. <sup>16</sup>James D. Watson Institute of Genome Sciences, Hangzhou, 310058, China. <sup>17</sup>Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming, 650201, China. Xian Qiao, Rui Su, Yang Wang and Ruijun Wang contributed equally to this work. Correspondence and requests for materials should be addressed to X.C. (email: [chenxiaoli@genomics.cn](mailto:chenxiaoli@genomics.cn)) or X.X. (email: [xuxun@genomics.cn](mailto:xuxun@genomics.cn)) or J.L. (email: [lijinquan\\_nd@126.com](mailto:lijinquan_nd@126.com))

breeding, which assumed that genes with related functions were often clustered in the genome to control biological traits<sup>3</sup>. Since then, the QTLs in association with genes or chromosome segments of interest have been widely studied to delineate complex animal traits<sup>4–6</sup>. In addition, marker-assisted selection (MAS)<sup>7</sup> was devised to introduce desirable QTLs into an animal population or increase the proportion of desirable QTLs in the gene pool. Despite its usefulness, the MAS program used a small number of DNA markers to trace limited numbers of QTLs<sup>8</sup>. This disadvantage led to the development of genomic selection (GS), which aimed to use simultaneously all available genome-wide dense SNP markers to predict breeding values<sup>9</sup>. Another strategy known as genome-wide association (GWA) analysis was also proposed, which believed that specific SNP markers could be in genetic linkage disequilibrium with a causative mutation affecting animal traits<sup>10</sup>. Therefore, identification of these significant genome-wide SNP markers is important for studying complex traits.

Thanks to the development of next-generation sequencing technology, the cost of large-scale genotyping has been reduced dramatically. This technological advancement provides the possibility of designing and utilizing high-throughput SNP chip for animal breeding<sup>11</sup>. Among all available products on the market, the GoldenGate and Infinium analyses are extensively used in animal genetic studies. Both assays are based on Illumina's Bead Chip technology, which involves direct hybridization of whole-genome amplified genomic DNA to a bead array of 50-mer locus-specific primers<sup>12–14</sup>, an enzymatic-based extension assay, a sandwich-based immunohistochemistry assay, and the final imaging by a two-color confocal laser system (Fig. 1a)<sup>15</sup>.

Target enrichment prior to sequencing is a useful method, in that specific portions of a genome can be analyzed to a greater depth<sup>16</sup>. This is due to the utilization of capture probes designed to select DNA regions of great interest. Compared with single locus genotyping, targeted sequencing can not only obtain a large scale of low density to high density SNPs, but also provide more information about the SNP variations, insertions/deletions, and copy number variations<sup>17</sup>. The major target enrichment strategies include molecular inversion probes<sup>18</sup>, SHS<sup>19</sup>, microarray-based GS<sup>20</sup>, and so on. Figure 1b shows the schematic steps of a SHS-based targeted sequencing process. Considering the sensitivity of genotyping, the uniform depth of coverage, and the scaling of reagent cost, SHS-based targeted sequencing is suitable for medium and large projects<sup>21</sup>.

Even though current commercial SNP chips based on the Bead Chip technology for different domestic animals have been successful<sup>22–27</sup>, we propose to design an animal SNP chip using the SHS-based target enrichment strategy for the first time. As an update to the international collaboration of goat research, we chose cashmere goat as our model animal. A 52 K SNP chip for goat (Illumina Inc., San Diego, CA) developed by the International Goat Genome Consortium included sequencing data from Saanen, Alpine, Creole, Boer, Katjang, and Savanna goat breeds<sup>28</sup>. This chip has been widely used to study the genetic diversity<sup>29</sup>, population structure<sup>30</sup>, effective population size<sup>31</sup>, QTL detection<sup>32</sup>, and GS in multiple goat populations<sup>33</sup>. No special SNP chip is designed for cashmere goat till now.

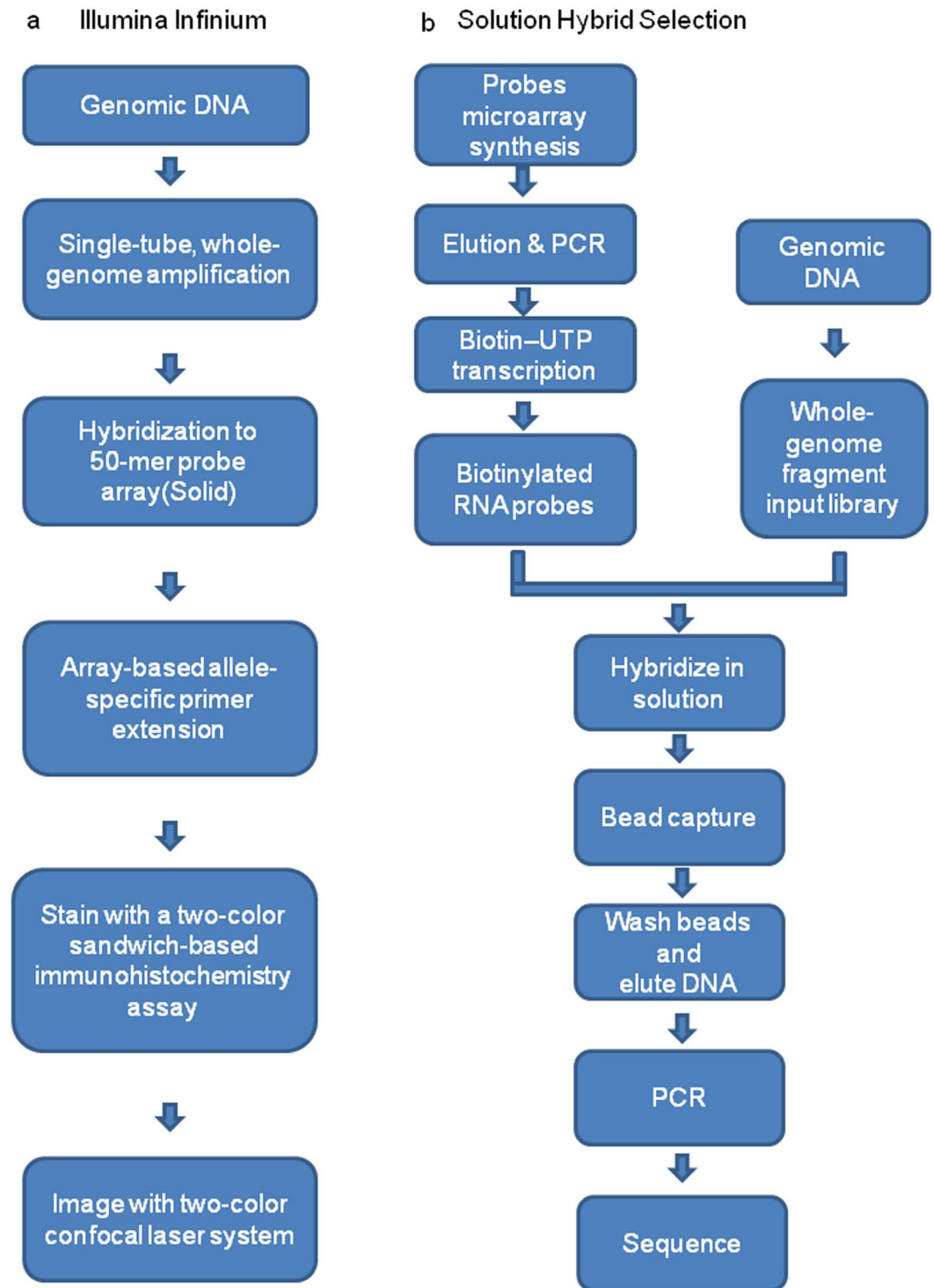
Cashmere goat is a multipurpose breed that adapts well to the desert and semi-desert pastoral environments. This goat breed produces high-quality cashmere fiber, which is crucial to the world textile industry. It is estimated that cashmere goat herding has contributed substantial economic benefits to the local people in the remote regions of developing Asian countries<sup>34, 35</sup>, and the downstream industries have increased international trade between Asia and the developed world<sup>36, 37</sup>. The meat from free-ranging cashmere goat is also considered as delicacy, and has aroused the interest of many meat markets. Here, a moderate-density SNP chip for cashmere goat was designed using the SHS-based target enrichment strategy. This chip was subsequently tested in a population of Inner Mongolia cashmere goats, through which several potential loci related to cashmere goat traits were obtained through GWA analysis.

## Results

**66 K SHS-based target enrichment SNP chip design for cashmere goat.** A total of 2,801,066 SNPs were called from the genome sequencing data of 73 cashmere goats (Fig. 2, Supplementary Table 1). These SNPs were used as candidates for SNP selection. After the first three steps of initial data filtering, 878,372 SNPs were obtained for the probe design process (Fig. 3a). The secondary filtering process of the designed probes yielded 64,898 SNPs from the cashmere goat population for the SNP chip. At this step, we decided to add 858 SNPs (courtesy of Dr. Jiang, data from 17 cashmere goats and 21 non-cashmere goats) in some genes related to wool fiber traits to our SNP pool. After the removal of redundant SNPs, a set of 65,620 SNPs were chosen for the chip design (Supplementary Table 2).

These SNPs spread over 30 chromosomes and 27 scaffolds (Fig. 3b). The frequency distribution of the spacing between adjacent SNPs revealed that about 95.7% of the SNPs were within 15 kb–70 kb to their closest neighbor counterparts (Fig. 3c), indicating no selection bias in the SNP-dense genomic regions. Additionally, the majority of the selected SNPs had a minor allele frequency (MAF) score greater than 0.2 (Fig. 3d). Based on these SNPs and their probe designs, a 188 K probe library was synthesized and separated in two 94 K sub-libraries. This marked the creation of a 66 K SHS-based target enrichment SNP chip for cashmere goat (Supplementary Table 3).

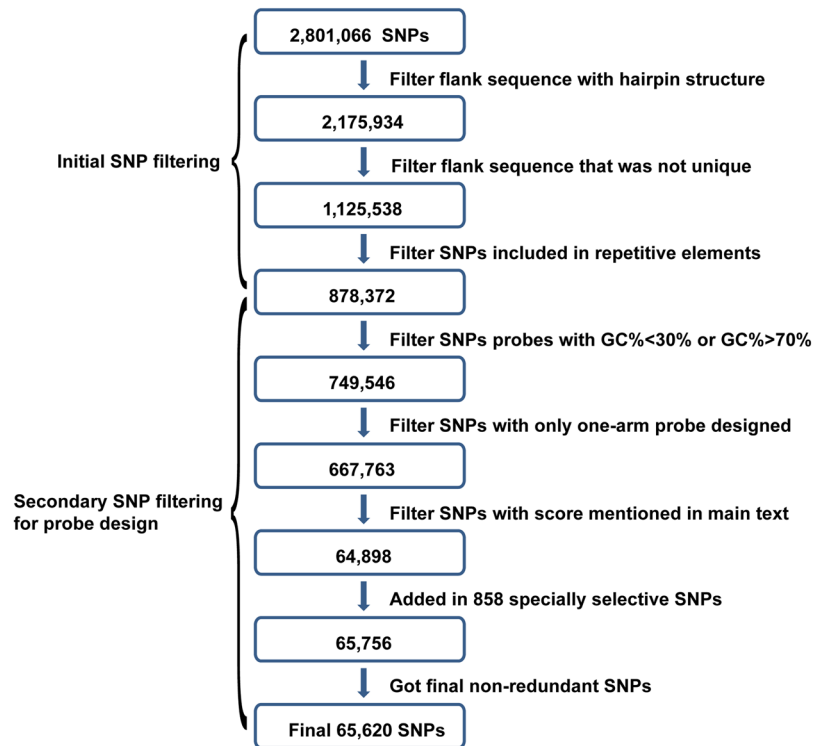
**Population of cashmere goats for GWA analysis.** The average measurements of the diameters of cashmere fibers from 1,438 cashmere goats were ranked from the smallest value to the largest value with a range from 12.9  $\mu\text{m}$  to 18.8  $\mu\text{m}$ . At the population level, the diameter of cashmere fiber conformed to a normal distribution with 22 intervals (Supplementary Figure, K-S test,  $P = 0.174$ ). We selected 300 cashmere goats from each of the two tails of this distribution, so that the average diameters of cashmere fibers had the largest difference between the two subpopulations (14.9  $\mu\text{m}$  vs. 16.2  $\mu\text{m}$ , Supplementary Table 4). After the removal of 164 cashmere goats without whole blood samples, a group of 436 cashmere goats (213 with smaller fiber diameter, 223 with bigger fiber diameter) were used for GWA analysis to test the 66 K SNP chip.



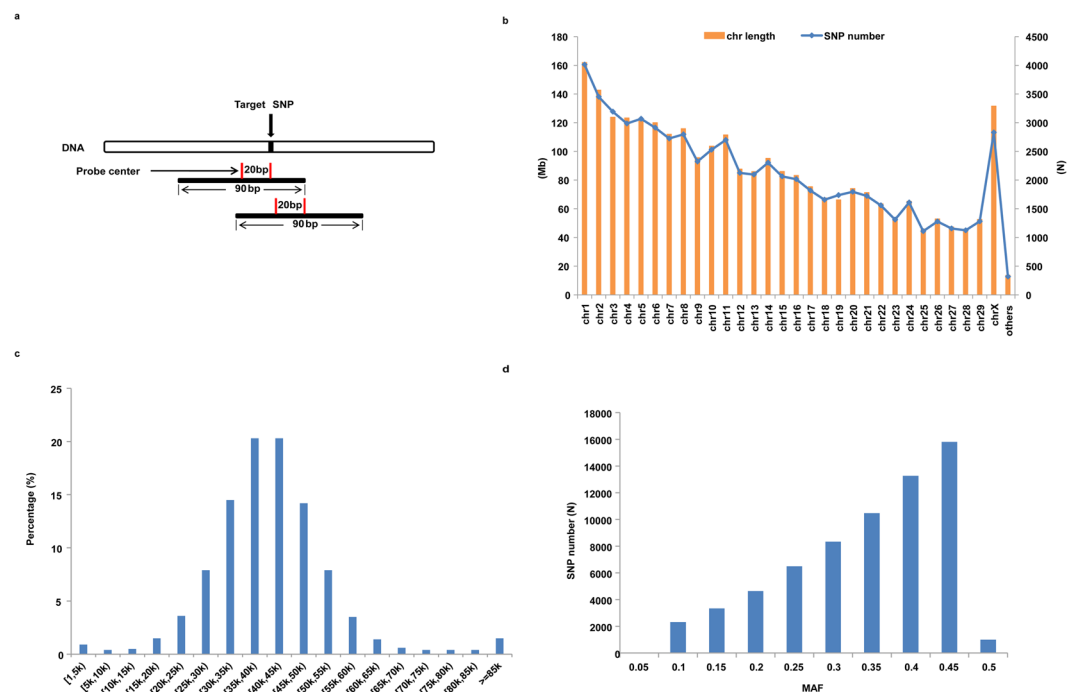
**Figure 1.** Schematic comparison between Illumina Infinium assay and solution hybrid selection (SHS)-based targeted sequencing method. **(a)** The schematic diagram of Illumina Infinium assay. **(b)** The schematic diagram of SHS-based targeted sequencing method.

**SNP detection from 436 cashmere goats using the 66 K SNP chip.** The 66 K SNP chip was used for the targeted sequencing of 436 cashmere goat genomes. About 15 Gb 50 bp single-end high-quality reads was produced for each animal on the BGISEQ-1000 platform. After initial data processing and population SNP calling, a total of ~407 K SNPs were obtained from these 436 animals. All sequencing reads covered about 95.3–99.8% of target SNPs on the chip depending on the individual goat.

In order to test the accuracy of SNP detection on different sequencing platforms, 11 target-enriched DNA sequencing libraries were randomly chosen from 436 samples and sequenced on an Hiseq, 2000 platform. These



**Figure 2.** SNP selection and probe design workflow for the 66 K SNP chip.



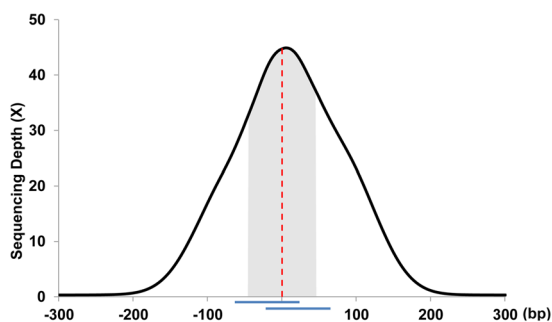
**Figure 3.** The characterization of final SNPs selected for designing the 66 K SNP chip. (a) The schematic graph of probe design. The distance between the center of probe and the target SNP is shown. (b) The number of SNPs corresponds to the lengths of scaffolds/chromosomes in the goat genome. (c) The frequency distribution of the spaces between neighboring SNPs. (d) The frequency distribution of the MAFs of the selected SNPs.

List	amount or percentage
Aggregate length of target flanking region	19 Mb
Aggregate length of baits	11 Mb
Total raw unfiltered sequence	2,100 Mb
Aligned to reference genome	1,849 Mb (87%)
Uniquely aligned to reference genome	1,666 Mb (78%)
Aligned to flanking region	1,149 Mb (56%)
Aligned to off target region	700 Mb (30%)
Target SNP coverage	95.3–99.8%

**Table 1.** Detailed description of sequences generated from catches.

List	before filter (fraction)
MAF (<0.05) in 65 K target SNP	109 (<0.2%)
Average sample missing in 65 K target SNP	5 (1.26%)
Depth in 65 K target SNP	~40X
Depth in 200 bp flanking region	~20X
Depth not in target capture region	~2X

**Table 2.** Detailed depth description of different type of position.



**Figure 4.** The sequencing depths of target SNP and its flanking regions. The curve is the base-by-base sequencing coverage around the target SNP. The probes are marked as the horizontal blue bars. The red dashed line in 0 position indicated the target SNP location.

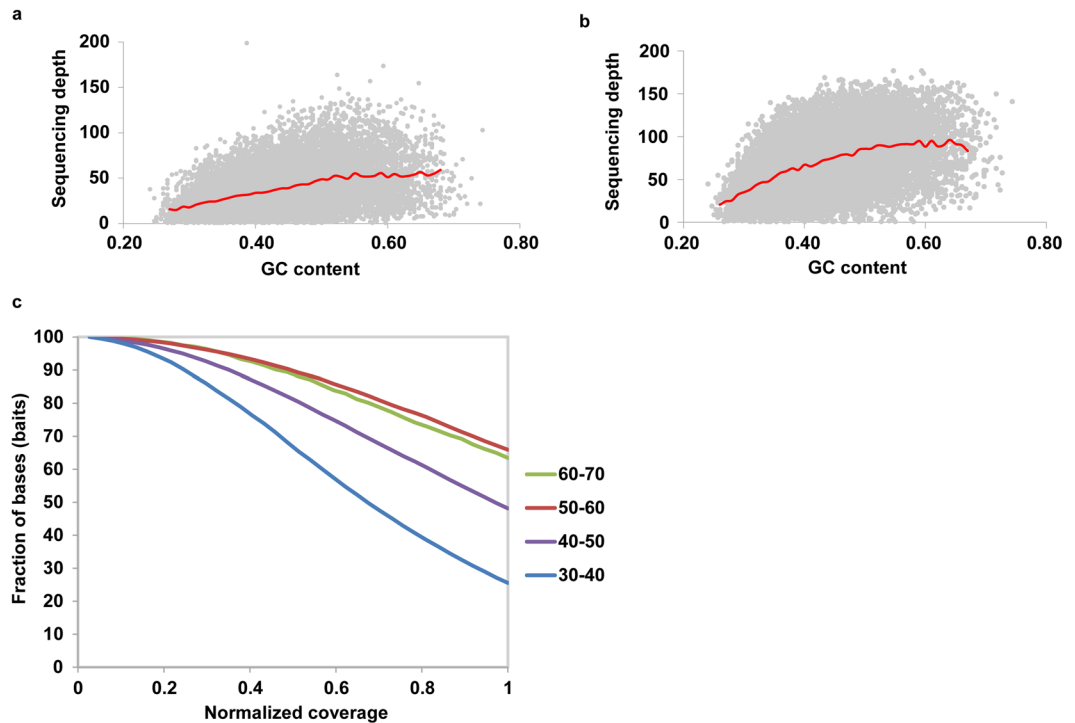
11 libraries produced 46,395 SNPs from BGISEQ-1000, and 50,063 SNPs from Hiseq. 2000, respectively. The number of shared SNPs was 45,601, which accounted for more than 90% of detected SNPs on both platforms. This suggests that the 66 K SNP chip may be suitable for different sequencing platforms.

The analysis showed that about 87% reads could be aligned to the reference genome, which included 56% reads within 300 bp from the target SNPs (Table 1). The estimation of sequencing depths of target SNPs and their 300 bp flanking regions showed a unimodal, center-weighted coverage distribution (Fig. 4). The majority of sequencing reads were located around target SNPs. The average sequencing depth of the target SNPs was 40X (Table 2). The average sequencing depth of the flanking regions decreased gradually as they distanced from the target SNPs. There were almost no reads covering regions 200 bp away from the target SNPs (depth ~2X). These results indicate that our 66 K SNP chip could effectively enrich the target SNP regions and narrow the flanking regions.

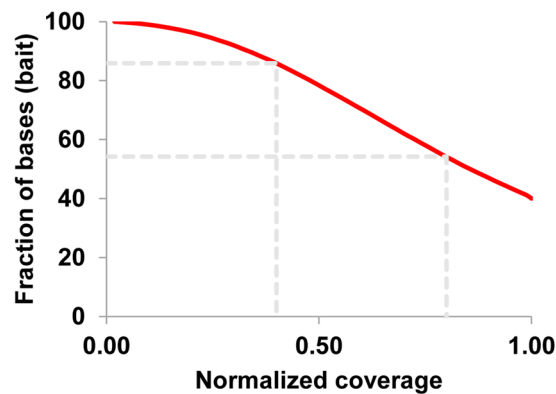
It has been reported that many factors during sequencing steps, including polymerase chain reaction<sup>38</sup>, size-selection<sup>39</sup>, and probability of sequencing-errors<sup>40</sup>, could cause GC-content bias. To inspect the effect of sequencing depths on GC bias, GC contents were compared against mean read depths across all capture regions, with an average depth of 37 and 69, respectively (Fig. 5a,b). In addition, the genomic regions with 50–70% GC content had higher coverage and higher depth. In comparison, those genomic regions with 30–50% GC content had relative lower coverage and depth (Fig. 5c).

Uniformity of coverage is another important parameter for targeted sequencing. The coverage of each target SNP was normalized to the mean coverage observed across the entire set (Fig. 6). Plot showed that 54% of the target SNPs had at least 80% of mean coverage, and 86% of the target SNPs had at least 40% of mean coverage.

**SNP selection for GWA analysis.** A total of 5,501,922 raw SNPs were obtained, 5,131,533 SNP with call rates <0.90. 13 animals with genotypes call rates <0.90, after removing non-biallelic SNPs (singleton/multi-allelic



**Figure 5.** The relationship between GC content and sequencing depths. (a) Average depth of target SNP was 37. (b) Average depth of target SNP was 69. The red lines indicated mean depth. (c) Normalized coverage distribution plots in four different GC% categories.



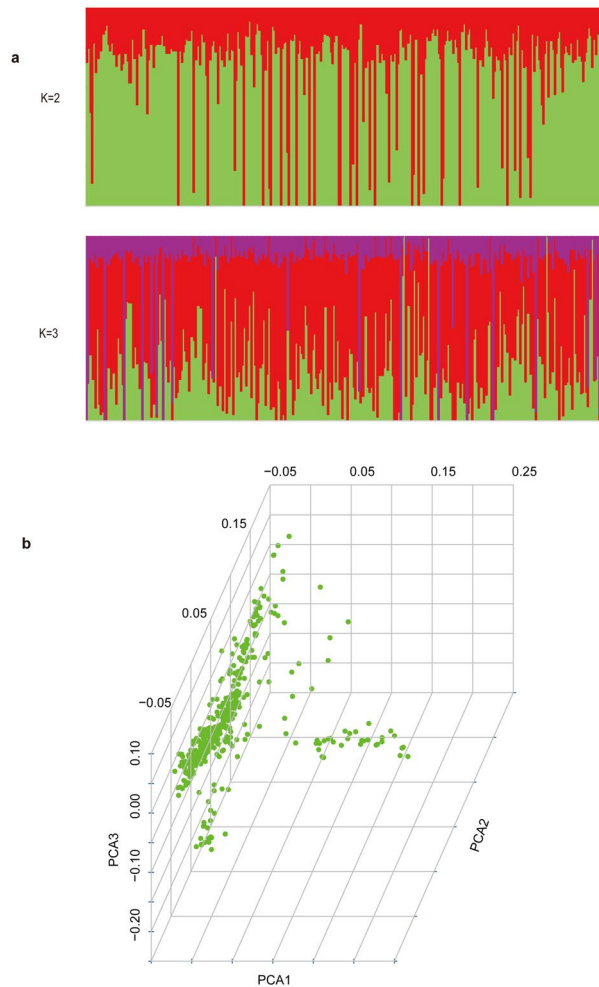
**Figure 6.** Normalized coverage distribution plots.

SNP), 423 cashmere goats with 370,389 SNP were left. The MAF and Hardy-Weinberg equilibrium (HWE) screening resulted in 161,125 SNPs for the following analyses. Among these, 55,863 SNPs matched the target SNP loci on the 66 K SNP chip.

**Population structure analysis.** We used PLINK software to conduct the population stratification analysis based on pairwise IBS distance, and found all the 423 samples were clumped into the same cluster. We ran STRUCTURE to determine the genetic ancestry constituents of samples and found that all samples almost have the same mixed ancestry when  $K = 2$  or  $K = 3$  (Fig. 7a). The results of principal component analysis (PCA) analysis showed that three principal components (PC1, PC2 and PC3) did not divide the populations (Fig. 7b).

**GWA analysis for cashmere fiber trait.** The cashmere fiber quality is a complex quantitative trait related to several parameters, including fiber length, guard hair length, fiber fineness (diameter), and combed cashmere fiber weight<sup>41</sup>. In this study, 423 goat individuals were subjected to quantitative trait association analysis. As shown in the Manhattan plot, no association were observed between SNPs and fiber fineness at the genome-wide level statistical significance ( $P$ -value  $< 4.2 \times 10^{-7}$ , Fig. 8a), However, several peaks were observed with marginal statistical significance. This result corresponded to the hypothesis that the genetic control of fiber fineness





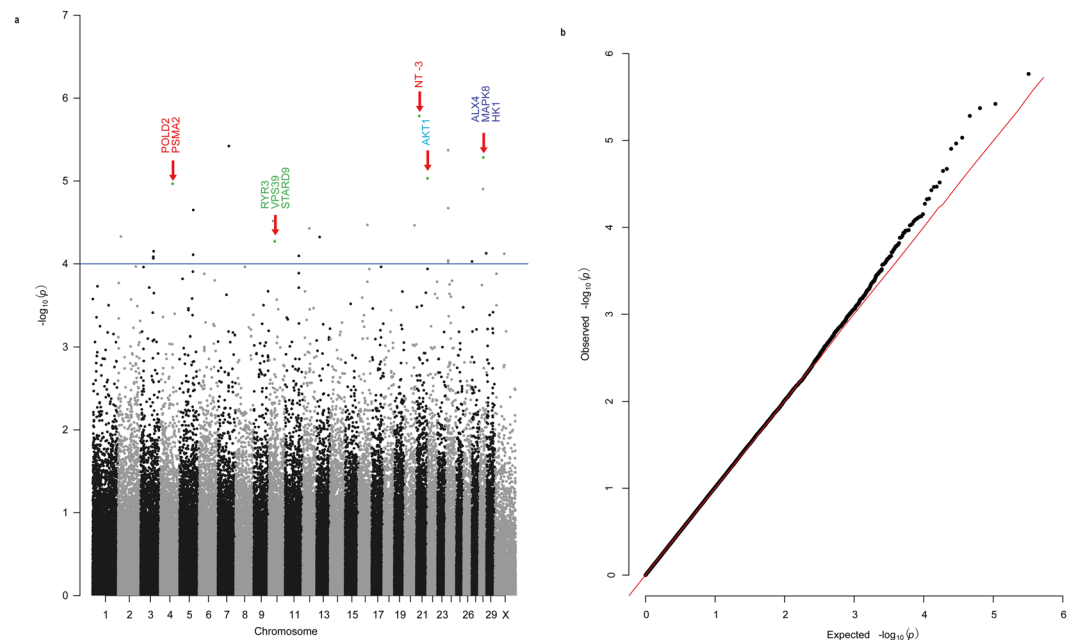
**Figure 7.** Population structure analysis of Inner Mongolia cashmere goat (Erlangshan). **(a)** Population structure analysis using STRUCTURE. Each sample is represented by a vertical bar. Every color represents one ancestral population and the length of each colored segment in each vertical bar represents the proportion contributed by ancestral populations. **(b)** PCA using all identified SNPs as markers. Each dot is an index to samples, most samples cluster together.

involves multiple QTLs of minor effects<sup>42</sup>. No genomic inflation was observed according to the quantile-quantile plots with  $\lambda$  value of 1.013 (Fig. 8b).

To extract some useful information from this analysis, 26 top-hit loci were chosen for the annotation against the KEGG database (<http://www.kegg.jp/>) (Supplementary Table 5). The result showed that genes with some top-hit SNPs are involved in signal pathways that are associated with the development of hair follicles (e.g. 4 enriched in MAPK, 2 in Wnt, 2 in TGF and 1 in Notch and so on)<sup>43</sup>. Some other genes with top-hit SNPs were reported to be important for skin or hair follicle growth and development (Table 3).

## Discussion

**SNP genotyping by target enrichment.** Compared with other genotyping methods, SNP genotyping by target enrichment is a cost-effective method. Take the 66 K SHS-based target enrichment SNP chip for cashmere goat as an example, the cost was estimated to be about 214 dollars per sample for an expected sequencing depth of 10X and 138 dollars per sample for sequencing depth of 6X. Despite the low price, this chip can capture regions up to 200 bp away from target SNPs. It also shows that the sequencing depth was highest for the target SNP, suggesting a high efficiency for target enrichment. This feature helps customize the amount of sequencing data according to user's need. For instance, the center of the SNP probes in this study was about 20 bp away from the target SNP. If a sequencing depth of 20X for target loci was expected, about 8 kb sequencing data would be sufficient for covering target regions. The required data would decrease to even lower amount, if the capture region is of a narrower scale. Another prominent advantage of this method is the flexibility showed in the chip design. The potential number of SNPs is not limited to a certain number. More SNP loci could be added to the chip for extra costs. This suggests that the target enrichment-based SNP chips can be designed for other species in the future, and be used for larger and more complex research projects.



**Figure 8.** Manhattan and quantile-quantile plot of GWAS for fiber fineness in cashmere goat. **(a)** Manhattan plot of the compressed MLM model for fiber fineness. **(b)** Quantile-quantile plot of the compressed MLM model for fiber fineness.

ID	Symbol name	P value	Signal Pathway	Type	Reference
K04456	AKT1	9.29E-06	MAPK&Notch	Hair follicle	Dipoï, N. <i>et al.</i> <sup>49</sup>
					Yang, Z. Z. <i>et al.</i> <sup>50</sup>
					Mauro, T. M. <i>et al.</i> <sup>51</sup>
K09451	ALX4	7.46E-05	WNT&MAPK	Hair follicle	Kayserili, H. <i>et al.</i> <sup>52</sup>
					Boras, K. <i>et al.</i> <sup>55</sup>
K00844	HK1	5.21E-06	TGF&Shh	Hair follicle	Ellis, T. S. I. <i>et al.</i> <sup>357</sup>
K05101	NT-3 growth factor receptor	1.72E-06		Hair follicle	Botchkarev, V. A. <i>et al.</i> <sup>58</sup>
K02328	POLD2	1.08E-05	WNT	Skin	Baldeck, N. <i>et al.</i> <sup>59</sup>
K02726	PSMA2	1.08E-05	WNT	Skin	Park <i>et al.</i> <sup>60</sup>
K04440	MAPK8 (JNKs)	7.46E-05	WNT&MAPK	Skin	Schumacher, M. <i>et al.</i> <sup>61</sup>
K04963	RYR3	3.04E-05	TGF&MAPK	Skin	Denda, S. <i>et al.</i> <sup>62</sup>
K20183	VPS39 (TLP)	5.36E-05	MAPK	Skin	Wang, X. <i>et al.</i> <sup>63</sup>
K16491	STARD9	5.36E-05	MAPK	Skin	Lin. <i>et al.</i> <sup>64</sup>

**Table 3.** Biological inferences analysis of top-hit SNP.

### Comparison of the 66 K SHS-based target enrichment SNP chip for cashmere goat with the 52 K SNP Bead Chip for goat.

The SNPs on the 52 K Bead Chip were identified from a total of 97 animals of six goat breeds, without cashmere goat. This chip was used to investigate the genetic diversities and populations structures of Italian goats and South African goats, respectively<sup>29,30</sup>. The result of the first study indicated that the genetic diversities of the present-day Italian goat populations were shaped by the combined effects of drift, gene flow, and recent demographic history<sup>29</sup>. The author of the second study, using a subset of 44,660 SNPs from 216 individuals, argued that the indigenous South African goats had a high genetic diversity<sup>30</sup>. The 66 K SHS-based target enrichment SNP chip, in comparison, is a suitable genomic research tool for cashmere goat. This chip was designed with extensive sample sources, especially Chinese cashmere goat breeds. It is therefore very useful for population genetic analysis.

Lashmar *et al.*<sup>44</sup> attempted to use the 52 K SNP Bead Chip to investigate genetic markers that are associated with fiber-producing traits in goat breeds. It is worth noting that only 10,659 out of the total 53,347 SNPs could be used for association analyses. In addition, both Becker *et al.*<sup>45</sup> and Martin *et al.*<sup>46</sup> used the 52 K chip to find SNPs associated with the coat color trait in the Coppernecked goats and Saanen goats, respectively. The results of all



of these studies were nevertheless not very ideal. In comparison, about 161,125 valid SNPs were captured from 436 cashmere goats with our 66 K chip for subsequent GWA analyses, showing a much potent way of acquiring large genetic information. Even though no statistically significant genetic contributors to cashmere fineness were identified in the GWA analyses, some top-hit SNPs seemed to be associated with physiological pathways that are important for hair development. One possible explanation for this result might be the moderate phenotypic variance in cashmere fineness (14.9 $\mu$ m vs. 16.2 $\mu$ m) in the cashmere goat populations. Another possible explanation might be that cashmere fineness is controlled by a combination of genetic and environmental factors.

It is now a common practice to use genome-wide SNPs generated by whole-genome sequencing or high-density SNP chip to carry out animal genetics research. It is not limited to traditional genomics research, but more and more used in studies on interesting animal traits. For example, Yang *et al.* showed the impact of global climate change on native sheep rapid adaptations to extreme environments. Kijas *et al.* and Lv *et al.* used Ovine SNP50K Beadchip to study the genetic history of global sheep breeds<sup>47</sup>, the adaptation to climate-mediated selection<sup>48</sup>, and so on. These successful application cases showed the great potential of genome-wide SNPs and high-density SNP chip technology. Even though our chip is most suitable for analyzing cashmere fiber trait, it is worth mentioning that it can be modified to study other cashmere goat traits, and used in animal breeding. Our chip design method can also be exploited for other species in the future.

**The top-hit SNPs associated with fineness.** The result of KEGG analysis discovered some top-hit SNPs near the genes, which are involved in hair follicle development. For example, *AKT1* gene belongs to the Notch and MAPK signal pathway (Table 3). It has been reported that *AKT1* activity is required for hair peg elongation during the hair follicle development. The hair follicles were distinctly reduced in size in *Akt<sup>+/-</sup>Akt3<sup>+/-</sup>* animals compared with their wild counterparts<sup>49–51</sup>. Another gene, *ALX4*, plays a critical role in skin and hair follicle development in human<sup>52</sup>. *ALX4* binds to *LEF-1*, a key regulatory factor for hair follicle development<sup>53, 54</sup>, and regulates its N-CAM promoter activity. Both *LEF-1* and *Alx4* knockout animals have defects in the hair follicle development<sup>55</sup>. In addition, *HK1* gene regulates the Shh pathway in the hair follicle<sup>56</sup> to inhibit the embryonic hair follicle morphogenesis<sup>57</sup>. Besides, the gene *NT-3* in the top-hit SNP locus (Fig. 8a) encodes a growth factor receptor, which is known to be important for hair follicle development<sup>58</sup>. Other genes around the top-hit SNPs are proved to be related of hair and skin<sup>59–64</sup>.

It can be concluded from the results that the breed-based application of our 66k goat SNP chip for cashmere fineness trait is possible. Even though no significant major singular contributor to fineness was included in our method, some top-hit SNPs proved to be important to hair follicle observed in this study by a medium density chip. It suggests that the 66k goat SNP chip will allow for applications such as GWAS, diversity studies, selection signatures and eventually genomic selection in the future.

## Methods

**Cashmere goat populations used in this study.** One group of 73 female cashmere goats sampled from four pastoral locations in Inner Mongolia and Liaoning Provinces (19 Erlangshan, 16 Alashan, 19 Albas, and 19 Liaoning) were used in the extraction of SNP variants and the subsequent design of a 66k SNP chip. A second group of 1,438 female goats from the Tonghetai Breeding Farm in Erlangshan were used in the testing of the designed SNP chip with GWA analysis (Supplementary Table 1). All goats used in this study were raised in the free-ranging style.

**Venous whole blood sample collection and genomic DNA extraction.** With the assistance of local herdsmen, trained veterinarians randomly chose three-year-old female cashmere goats from the populations, and collected 5 ml whole blood from the left jugular vein of each animal into a plastic collection tube containing 4% (w/v) sodium citrate. The blood samples were snap frozen in liquid nitrogen, and stored at  $-80^{\circ}\text{C}$  until further processing. Genomic DNA was extracted from whole blood samples with the AXYGEM Blood and Tissue Extraction Kit (Corning, USA) according to the manufacturer's instructions. The extracted DNA was subjected to electrophoresis in 2% agarose gel and stained with ethidium bromide to assess overall quality. The DNA concentration was determined by Quant-iT<sup>TM</sup> PicoGreen<sup>®</sup> dsDNA Reagent and Kits (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. All animal procedures were approved by the Inner Mongolia Agriculture University Animal Care and Use Committee in accordance with the National Animal Care Standard (GB 14925–2001). All experiments were performed in accordance with relevant guidelines and regulations. All efforts were made to minimize animal suffering.

**Chip design: library construction, sequencing, SNP discovery and characterization.** For the group of 73 cashmere goats, paired-end libraries were constructed for each individual animal with an insert size of 300 bp from  $\sim 2\mu\text{g}$  of sheared genomic DNA according to the procedures of NEB DNA Library Prep Kit for Illumina (NEB, USA). These libraries were sequenced on an Illumina HiSeq. 4000 platform (Illumina; CA, USA) using a PE-100 module. After data filtering, high quality reads were mapped to the goat reference genome (version 2.0)<sup>65</sup> using the Burrows-Wheeler Aligner (version 0.7.10-r789)<sup>66</sup> with default settings. The software SAMtools<sup>67</sup> was used to convert file format from SAM to BAM. The package Picard (<http://broadinstitute.github.io/picard/>) was used to sort BAM files by coordinate and mark PCR duplications.

After the BWA alignment process, 'RealignerTargetCreator' and 'IndelRealigner' in the Genome Analysis Toolkit (GATK, version 3.3–0-g37228af)<sup>68</sup> were used to obtain duplication-free reads with default settings. Next, 'HaplotypeCaller' in the GATK was used to generate a single call set from the sequencing data of all 73 individuals by joint calling with the parameter '-stand\_call\_conf 30 -stand\_emit\_conf 10'. The SNP data were then distinguished from the InDel data by 'selectVariants' in the GATK. The criteria<sup>68</sup> used to exclude false positive SNP data were the following: (a) hard filtration with the parameter 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum

$< -12.5$  ||  $\text{ReadPosRankSum} < -8.0$ ; (b) total depth does not range from 80 to 1000; (c) missing data rate  $>20\%$  and  $>20\%$  individual depth  $<2$ .

**Chip design: SNP selection and probe design.** In brief, the probe design process includes the selection of SNP-containing genomic regions and the optimization of the probe hybridization efficiency (Fig. 2). The goal is to cover the whole genome as much as possible while achieving a low off-target hybridization rate. To this end, SNPs were initially filtered according to the following criteria: (1) the flanking sequences within 150 bp on either side of a SNP should not form any  $>8$  bp hairpin structure; (2) the flanking sequences on either side of a SNP should be unique; (3) Exclude any SNPs in the repetitive elements. The resultant SNPs were viewed as candidates for designing SNP probes.

A SNP probe is a 120 bp oligonucleotide, which contains a 90 bp target-specific bait sequence and two 15 bp PCR primer ends (5'-GAA GCG AGG ATC AAC [N90] CAT TGC GTG AAC CGA-3'). The 90 bp target-specific bait sequence pairs with the genomic region containing the target SNP. In this study, a series of probes were designed to cover each specific SNP with the stipulation that the center of the probe should be less than 50 bp away from the target SNP. These probes were screened by the GC content criterion (between 30% and 70%). SNPs with no viable probe designs or with only one viable probe design were removed from the candidate pool. If multiple probe designs were available for a SNP, two or three best probes were selected on the condition that the center of the probe should be about 20 bp away from the target SNP (Fig. 3a). For some SNPs, if one of the only two viable probe designs contains any  $>8$  bp hairpin structure, the other probe will be kept and used twice the amount for the final chip. In addition, any probes that cover two SNPs will also be used twice the amount for the final chip.

In order to cover the goat genome on a 60–70 K SNP chip, the average interval between SNPs is estimated to be about 40 kb. We further optimized the target SNPs density in consecutive 40 kb genomic windows. A ranking score was calculated for each of the remaining SNPs according to the following formula<sup>69</sup>:

$$\text{Score} = \text{MAF} * (E - S) * \left( 1 - \frac{|a - (E - S)/2|}{(E - S)/2} \right)$$

MAF is the minor allele frequency.  $a$  is the coordinate of a SNP.  $S$  and  $E$  are the initial coordinate, and stop coordinate of the 40 kb genomic window, respectively. In SNP-dense genomic regions, only those with high ranking scores were kept for the chip design.

**Chip design: SNP probe synthesis, processing, and the final chip.** All final SNP probes were obtained with a CustomArray B3™ Synthesizer (CustomArray, Washington DC, USA) according to the manufacturer's instructions. The probe libraries were aminolyzed, purified, and then dissolved in  $10\times$  TE buffer (pH = 8.0). The probe libraries were amplified by PCR with primers A (5'-GAA GCG AGG ATC AAC-3') and B (5'-TCG GTT CAC GCA ATG-3'). With the addition of SP6 promoter sequences, the probe libraries were transcribed into RNA bait libraries with SP6 RNA polymerase. They were then labeled by biotin. After purification and quality control, biotinylated RNA bait libraries were prepared for hybridization and stored at  $-70^\circ\text{C}$  (Fig. 1b), which were the final product of the SHS-based target enrichment SNP chip.

**Cashmere fiber sample collection and analysis.** For the group of 1,000 cashmere goats for GWA analysis, about one gram of cashmere fiber sample was obtained from the left scapular region of each individual animal. The average diameter of the cashmere fibers from each individual was assessed by an OFDA-2000 optical-based fiber diameter analyzer (BSC Electronics, Australia) according to the manufacturer's instructions.

**Targeted sequencing of 436 cashmere goats.** 1  $\mu\text{g}$  high-quality genomic DNA from each cashmere goat was subjected to sonication. The DNA fragments of 150–250 bp in lengths were selected for the targeted sequencing. These fragments were end-repaired before being ligated to the Ad153 $\Omega$ \_2B adapter (BGI Shenzhen, China, unpublished). They were then amplified by a round of standard PCR. The biotinylated RNA bait libraries (see methods above) were used to capture and enrich SNP-containing DNA fragments. Captured DNA fragments were subjected to a round of standard PCR, and the PCR products were circularized and made ready for sequencing (Fig. 1b) on a BGISEQ-1000 platform (BGI Shenzhen, China) with a SE50 module.

**Variant detection in the GWA population.** Raw sequencing reads were filtered according to the following criteria: (1) if a read has  $>10$  percent of bases as N; (2) if a read has  $>40$  percent of low-quality (value  $\leq 10$ ) bases; (3) if a read is contaminated by the adaptor sequence or produced by PCR duplication. The resultant clean reads were then mapped to the goat reference genome (v2.0) using BWA with parameters “-m 200000 -l 20 -k 2 -t 30”. The results were transformed into indexed BAM files with SAMtools (version 0.1.18). Picard package (version 1.105) was used to remove duplicate reads. Reads coverage and depth were calculated from BAM files with “samtools depth”. The variants were called using the Genome Analysis Toolkit's HaplotypeCaller. After separating SNPs from Indel variants, SNPs was further filtered using the VariantFiltration package in GATK with parameters “-filterExpression “QD  $< 4.0$  || MQ  $< 40.0$  || ReadPosRankSum  $< -8.0$  || FS  $> 60.0$  || HaplotypeScore  $> 13.0$  || MQRankSum  $< -12.5$ ” -filterName LowQualFilter”.

**SNP data quality control.** The following quality control process for our data was conducted by Plink v1.07 (<http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml>) unless stated otherwise. Chromosomal variant cell format files were transformed into Plink format by VCFtools v0.1.13, during which the non-biallelic SNPs were automatically filtered out in the PED/MAP files. Because human ( $n = 23$ ) was set as the default chromosome handling type as in Plink, the parameter *-dog* ( $n = 39$ ) was added in front of each command line to ensure that

the system could process all goat chromosomes ( $n = 30$ ). SNPs with call rates  $< 0.90$  and samples with genotyping call rates  $< 0.90$  were removed for the further statistical analyses. The SNPs from the final goat individuals were subjected to quality control according to the following two criteria: (1) Remove SNPs with a very low MAF filtration ( $MAF < 0.01$ ); (2) Remove SNPs with significant deviations from HWE filtration ( $HWE < 0.001$ ).

**Population structure analysis.** Population substructure was investigated using Clustering, STRUCTURE<sup>70</sup> and PCA<sup>71</sup> based on using genomic SNPs. We used Plink to do stratification analysis based on pairwise identity-by-state (IBS) distance with option `-cluster-mc 2-ppc 0.05`. Further STRUCTURE was used to infer genetic ancestry constituents and assign individuals to subpopulation. We also performed a PCA following the procedure as reported<sup>71</sup>. The eigenvector decomposition using the R function `eigen`, and the significance of the eigenvectors was determined with a Tracey-Widom test.

**Association analysis.** The compressed mixed linear model (MLM) were used to identify association signals with the software EMMAX<sup>72</sup>. The basic model underlying this software can be written as

$$y_i = \beta_0 + \sum_{k=1}^M \beta_k X_{ik} + \varepsilon_i \quad (1)$$

$$\text{Var}(Y) = 2\delta_{ak}^2 \phi_k + 2\phi_a^2 + \delta_e^2 I \quad (2)$$

In equation (1) the vector  $Y = \{y_1, \dots, y_n\}$  contain the phenotypes of the individuals,  $\text{Var}(\varepsilon) = \delta_e^2 I$ ,  $\text{Var}(Y)$  was used to investigate the contribution of locus  $k$  to the phenotype which the effect of the genotype at locus  $k$  can be modeled as a main effect, whereas the relationships among all individuals are taken into account by means of variance components of random polygenic effects. We calculated an identity-by-state kinship matrix using the Affymetrix genotypes in EMMAX with command `"emmax-kin -v -h -s -d 10"`, pairwise relatedness matrix was used to represent the sample structure. Using a variance component model, we got an estimated covariance matrix that models the effect of genetic relatedness on the phenotypes. Animal pasture information was used as covariate matrix. For cashmere trait, the threshold  $P$ -value for declaring genome-wide significance ( $P < 4.2 \times 10^{-7}$ ) which was set to control genome-wide type 1 error rate. Manhattan plot was drawn by *qqman* package of R (v3.2.0). A 500 kb region on each side of peak SNP was searched for gene annotation.

## References

- Lush, J. L. Family Merit and Individual Merit as Bases for Selection. Part I. *The American Naturalist* **81**, 241–261 (1947).
- Schaeffer, L. R. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* **123**, 218–223 (2006).
- Geldermann, H. Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *Theoretical and Applied Genetics* **46**, 319–330 (1975).
- Soller, M. The use of loci associated with quantitative effects in dairy cattle improvement. *Animal Production* **27** (1978).
- Smith, C. & Simpson, S. P. The use of genetic polymorphisms in livestock improvement. *Journal of Animal Breeding and Genetics* **103**, 205–217 (1986).
- Dekkers, J. C. & Hospital, F. The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**, 22–32 (2002).
- Lande, R. & Thompson, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743 (1990).
- Golding, B., Hayes, B. & Goddard, M. Genome-wide association and genomic selection in animal breeding. *Genome* **53**, 876–883 (2010).
- Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
- Visscher, P. M. *et al.* Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).
- Hui Zhang, Z. W., Wang, S. and Hui, Li. Progress of genome wide association study in domestic animals. *Journal of Animal Science and Biotechnology* **3** (2012).
- Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics* **37**, 549 (2005).
- Steemers, F. J. & Gunderson, K. L. Illumina, Inc. *Pharmacogenomics* **6**, 777 (2005).
- Gunderson, K. L. *et al.* Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics* **7**, 641–648 (2006).
- Steemers, F. J. & Gunderson, K. L. Whole genome genotyping technologies on the BeadArray™ platform. *Biotechnology Journal* **2**, 41–49 (2007).
- Mertes, F. *et al.* Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics* **10**, 374 (2011).
- Li, W. *et al.* Identifying Human Genome-Wide CNV, LOH and UPD by Targeted Sequencing of Selected Regions. *PLoS ONE* **10**, e0123081 (2015).
- Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
- Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* **27**, 182 (2009).
- Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
- Teer, J. K. *et al.* Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Research* **20**, 1420 (2010).
- Georges, M. *et al.* A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. *PLoS Genetics* **8**, e1002451 (2012).
- Groenen, M. A. M. *et al.* The development and characterization of a 60K SNP chip for chicken. *BMC Genomics* **12** (2011).
- Orban, L. *et al.* Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLoS ONE* **4**, e6524 (2009).
- Lühken, G. Genetic testing for phenotype-causing variants in sheep and goats. *Molecular and Cellular Probes* **26**, 231–237 (2012).

26. Harris, B. L., Creagh, F. E., Winkelman, A. M. & Johnson, D. L. Experiences with the Illumina High Density Bovine BeadChip. *44* (2011).
27. Wiggans, G. R., Cooper, T. A., Van Tassell, C. P., Sonstegard, T. S. & Simpson, E. B. Technical note: Characteristics and use of the Illumina BovineLD and GeneSeek Genomic Profiler low-density bead chips for genomic evaluation I. *Journal of Dairy Science* **96**, 1258–1263 (2013).
28. Liu, Z. *et al.* Design and Characterization of a 52K SNP Chip for Goats. *PLoS ONE* **9**, e86227 (2014).
29. Nicoloso, L. *et al.* Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. *Genetics Selection Evolution* **47** (2015).
30. Mdladla, K., Dzomba, E. F., Huson, H. J. & Muchadeyi, F. C. Population genomic structure and linkage disequilibrium analysis of South African goat breeds using genome-wide SNP data. *Animal Genetics* **47**, 471–482 (2016).
31. Negrini, R. I E Past Population Size Changes of Italian Goat Breeds. *Plant & Animal Genome* (2014).
32. Palhira, I., Larroque, H., Virginie, C., Tosser-Klopp, G. & Rachel, R. Genetic Parameters and QTL Detection for Milking Speed in Dairy Alpine and Saanen Goats. *World Congress on Genetics Applied To Livestock Production* (2014).
33. Carillier, C. *et al.* A first step toward genomic selection in the multi-breed French dairy goat population. *Journal of Dairy Science* **96**, 7294–7305 (2013).
34. Lecraw, D., Eddleston, P. & McMahon, A. A Value Chain Analysis of the Mongolia Cashmere Industry. *Report prepared for USAID's Accelerating Sustainable Agriculture Program* (2005).
35. de Weijer, F. Cashmere Value Chain Analysis Afghanistan. Report prepared for USAID's Accelerating Sustainable Agriculture Program (2007).
36. McGregor, B. A. Australian cashmere: attributes and processing. *Rural Industries Research and Development Corporation* (2002).
37. Wang, Z. *et al.* Estimation of genetic parameters for fleece traits in yearling Inner Mongolia Cashmere goats. *Small Ruminant Research* **109**, 15–21 (2013).
38. Daniel, A. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, 1–14 (2011).
39. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005 (2008).
40. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39**, e90 (2011).
41. Zhang, Y. *et al.* Estimates of genetic parameters and genetic changes for fleece traits in Inner Mongolia cashmere goats. *Small Ruminant Research* **117**, 41–46 (2014).
42. Goodale, H. D. Dominant vs. Non-Dominant Genes: In the Multiple Factor Hypothesis of Size Inheritance. *Journal of Heredity* (1932).
43. Wei, J. *et al.* The transcriptome research progresses of skin hair follicle development. *Hereditas* **37**, 528–534 (2015).
44. Lashmar, S. F., Visser, C. & Van Marle-Köster, E. Validation of the 50k Illumina goat SNP chip in the South African Angora goat. *South African Journal of Animal Science* **45**, 56 (2015).
45. Becker, D. *et al.* The brown coat colour of Coppernecked goats is associated with a non-synonymous variant at the TYRP1 locus on chromosome 8. *Animal Genetics* **46**, 50–54 (2015).
46. Martin, P. M., Palhière, I., Ricard, A., TosserKlopp, G. & Rupp, R. Genome Wide Association Study Identifies New Loci Associated with Undesired Coat Color Phenotypes in Saanen Goats. *PLoS ONE* **11**, e0152426 (2016).
47. JW, K. *et al.* Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biology* **10**, e1001258 (2012).
48. Lv, F. H. *et al.* Adaptations to climate-mediated selective pressures in sheep. *Molecular Biology & Evolution* **31**, 3324 (2014).
49. Dipoi, N. *et al.* Epithelium-mesenchyme interactions control the activity of peroxisome proliferator-activated receptor beta/delta during hair follicle development. *Molecular & Cellular Biology* **25**, 1696–1712 (2005).
50. Yang, Z. Z. *et al.* Dosage-Dependent Effects of Akt1/Protein Kinase B $\alpha$  (PKB $\alpha$ ) and Akt3/PKB $\gamma$  on Thymus, Skin, and Cardiovascular and Nervous System Development in Mice. *Molecular & Cellular Biology* **25**, 10407 (2005).
51. Mauro, T. M. *et al.* Akt2 and SGK3 are both determinants of postnatal hair follicle development. *Faseb Journal Official Publication of the Federation of American Societies for Experimental Biology* **23**, 3193 (2009).
52. Kayserili, H. *et al.* ALX4 dysfunction disrupts craniofacial and epidermal development. *Human Molecular Genetics* **18**, 4357 (2009).
53. Kratochwil, K., Dull, M., Farinas, I., Galceran, J. & Grosschedl, R. Lef1 expression is activated by BMP-4 and regulates inductive tissue interactions in tooth and hair development. *Genes & Development* **10**, 1382–1394 (1996).
54. Petersson, M. *et al.* TCF/Lef1 activity controls establishment of diverse stem and progenitor cell compartments in mouse epidermis. *Embo Journal* **30**, 3004–3018 (2011).
55. Boras, K. & Hamel, P. A. Alx4 binding to LEF-1 regulates N-CAM promoter activity. *Journal of Biological Chemistry* **277**, 1120–1127 (2002).
56. Gallego, M. I., Beachy, P. A., Hennighausen, L. & Robinson, G. W. Differential requirements for shh in mammary tissue and hair follicle morphogenesis. *Developmental Biology* **249**, 131–139 (2002).
57. Ellis, T. S. I. *et al.* Overexpression of Sonic Hedgehog suppresses embryonic hair follicle morphogenesis. *Developmental Biology* **263**, 203–215 (2003).
58. Botchkarev, V. A. *et al.* A New Role for Neurotrophin-3: Involvement in the Regulation of Hair Follicle Regression (Catagen). *American Journal of Pathology* **153**, 785–799 (1998).
59. Baldeck, N. *et al.* FF483–484 motif of human Pol $\eta$  mediates its interaction with the POLD2 subunit of Pol $\delta$  and contributes to DNA damage tolerance. *Nucleic Acids Research* **43**, 2116–2125 (2015).
60. Park, D., Jeong, H. O., Kim, B. C., Ha, Y. M. & Chung, H. Y. Computational Approach to Identify Enzymes That Are Potential Therapeutic Candidates for Psoriasis. *Enzyme Research* **2011**, 826784 (2011).
61. Schumacher, M. *et al.* Efficient keratinocyte differentiation strictly depends on JNK-induced soluble factors in fibroblasts. *Journal of Investigative Dermatology* **134**, 1332 (2014).
62. Denda, S. *et al.* Ryanodine receptors are expressed in epidermal keratinocytes and associated with keratinocyte differentiation and epidermal permeability barrier homeostasis. *Journal of Investigative Dermatology* **132**, 69 (2012).
63. Wang, X. *et al.* Effects of TRAP-1-Like Protein (TLP) Gene on Collagen Synthesis Induced by TGF- $\beta$ /Smad Signaling in Human Dermal Fibroblasts. *PLoS ONE* **8**, e55899 (2013).
64. Lin, C. E., Kaptein, J. S. & Sheikh, J. Differential expression of microRNAs and their possible roles in patients with chronic idiopathic urticaria and active hives. *Allergy & Rhinology* **8**, 67 (2017).
65. Du, X. *et al.* An update of the goat genome assembly using dense radiation hybrid maps allows detailed analysis of evolutionary rearrangements in Bovidae. *BMC Genomics* **15**, 625 (2014).
66. Wu, Y. P. *et al.* A fine map for maternal lineage analysis by mitochondrial hypervariable region in 12 Chinese goat breeds. *Animal Science Journal* **80**, 372–380 (2009).
67. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
68. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).
69. Matukumalli, L. K. *et al.* Development and characterization of a high density SNP genotyping assay for cattle. *PloS one* **4**, e5350 (2009).



70. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
71. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genet* **2**, e190 (2006).
72. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**, 348–354 (2010).

### Acknowledgements

This work was financially supported by National high technology research and development program of China(863plan) (2013AA102506), National Key Basic Research Special Foundation of China(973plan) (2013CB835200), National Natural Science Foundation of China(31660639), and China Agriculture Research System (CARS-40-05).

### Author Contributions

X.C., W.W., X.X., Y.D. and J.L. designed the study. W.Z., Y.Z., Z.L., Z.W., H.J., R.D., Q.Z., T.Z., Y.J., Y.Z., R.W. and H.Z. did samples collection. Y.J. provided 858 SNPs in some genes related to wool fiber traits to our SNP pool. X.Q., Y.F., X.L., W.W. and B.L. did the sequencing experiment. Y.W., T.Y., S.H., Q.X., X.L., J.C., Y.Z., D.C., D.F., X.W., D.M. analyzed the data and design the chip. X.C., X.Q. and Y.T. participated in manuscript revision. R.S., R.W., W.C. conceived the overall study and revised the manuscript. All authors read and approved the final manuscript.

### Additional Information

**Supplementary information** accompanies this paper at doi:[10.1038/s41598-017-09285-z](https://doi.org/10.1038/s41598-017-09285-z)

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017