

# SCIENTIFIC REPORTS



OPEN

## Genome-wide analysis reveals differential selection involved with copy number variation in diverse Chinese Cattle

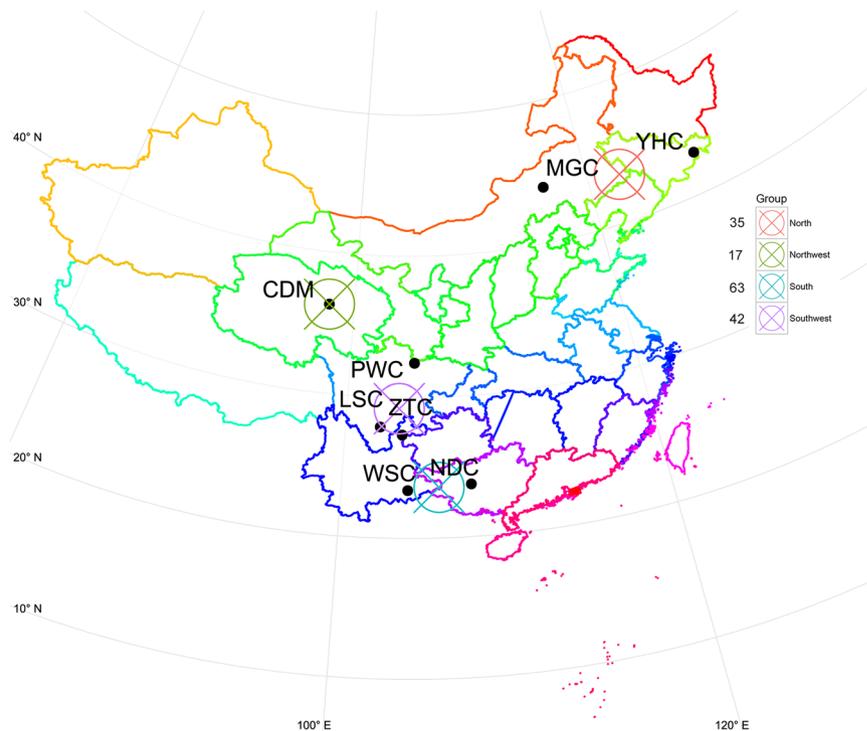
Liu Yang<sup>1,2</sup>, Lingyang Xu<sup>1</sup>, Bo Zhu<sup>1</sup>, Hong Niu<sup>1</sup>, Wengang Zhang<sup>1</sup>, Jian Miao<sup>1,3</sup>, Xiping Shi<sup>1,4</sup>, Ming Zhang<sup>2</sup>, Yan Chen<sup>1</sup>, Lupei Zhang<sup>1</sup>, Xue Gao<sup>1</sup>, Huijiang Gao<sup>1</sup>, Li Li<sup>2</sup>, George E. Liu<sup>5</sup> & Junya Li<sup>1</sup>

Copy number variations (CNVs) are defined as deletions, insertions, and duplications between two individuals of a species. To investigate the diversity and population-genetic properties of CNVs and their diverse selection patterns, we performed a genome-wide CNV analysis using high density SNP array in Chinese native cattle. In this study, we detected a total of 13,225 CNV events and 3,356 CNV regions (CNVRs), overlapping with 1,522 annotated genes. Among them, approximately 71.43 Mb of novel CNVRs were detected in the Chinese cattle population for the first time, representing the unique genomic resources in cattle. A new  $V_i$  statistic was proposed to estimate the region-specific divergence in CNVR for each group based on unbiased estimates of pairwise  $V_{57}$ . We obtained 12 and 62 candidate CNVRs at the top 1% and top 5% of genome-wide  $V_i$  value thresholds for each of four groups (North, Northwest, Southwest and South). Moreover, we identified many lineage-differentiated CNV genes across four groups, which were associated with several important molecular functions and biological processes, including metabolic process, response to stimulus, immune system, and others. Our findings provide some insights into understanding lineage-differentiated CNVs under divergent selection in the Chinese native cattle.

Copy number variations (CNVs) are defined as deletions, insertions, and duplications ranging from 50 base pairs (bp) to 5 million base pairs (Mbp) of genomic sequence between two individuals of a species<sup>1–4</sup>. Previous studies suggest CNV have potentially larger effects than other variation such as SNP, including changing gene structure and dosage, altering gene regulation and exposing recessive alleles<sup>5–7</sup>. CNV discovery studies have been extensively reported in human<sup>8</sup>, primates<sup>9</sup>, mouse<sup>10–12</sup>, zebrafish<sup>13</sup>, dog<sup>14–16</sup>, and livestock, including chicken<sup>17,18</sup>, pig<sup>19,20</sup>, sheep<sup>21–23</sup>, goat<sup>24</sup> and cattle<sup>25–29</sup>.

In recent years, many studies have revealed genomic diversity could be generated by the differential selection of CNVs in specific environments for adaptations<sup>30–33</sup>. In human, positive selection for a higher *AMY1* copy number enables better digestion of starchy foods<sup>34</sup>. A change in *CCL3L1* copy number is associated with markedly enhanced HIV/acquired immunodeficiency syndrome (AIDS) susceptibility<sup>35</sup>. The human *UGT2B17* gene shows significant copy-number diversity, and displays region-specific differences for metabolism in multiple populations<sup>36</sup>. Also, olfactory receptor (OR) genes with variable copy numbers among most mammalian species were found associate with population-specific differences in smell<sup>37</sup>. CNVs are specifically enriched among evolutionary “young” ORs, implying that CNVs may play an essential role in the origin of a gene or the emergence of new OR gene clusters<sup>38</sup>.

<sup>1</sup>Laboratory of Molecular Biology and Bovine Breeding, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing, 100193, China. <sup>2</sup>Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, Sichuan, 611130, China. <sup>3</sup>College of Animal Sciences, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China. <sup>4</sup>College of Animal Science and Technology, Agricultural University of Hebei, Baoding, Hebei, 071001, China. <sup>5</sup>Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, Maryland, 20705, USA. Liu Yang and Lingyang Xu contributed equally to this work. Correspondence and requests for materials should be addressed to L.X. (email: [xulingyang@163.com](mailto:xulingyang@163.com)) or J.L. (email: [jl1@iascaas.net.cn](mailto:jl1@iascaas.net.cn))



**Figure 1.** Geographic distribution of four groups from eight populations in China. YHC, MGC, CDM, PWC, LSC, ZTC, WSC and NDC are abbreviations for Yanhuang, Monggu, Caidamu, Pingwu, Liangshan, Zhaotong, Wenshan and Nandan, respectively. The numbers of individual sample for each group were shown beside figure legend. The distribution of cattle was summarized and visualized with R packages maps, mapproj, maptools and ggplot 2.

Domesticated cattle are one of the most economically important farm animals. The exploration of genetic diversity, conservation, selection and evolution of genomic variants in cattle have attracted much attention in past decades<sup>39</sup>. Specifically, investigations of population genetic properties and selection patterns involved with CNVs have gradually become an emerging research topic for farm animals. For instance, Xu *et al.* have investigated the population-genetic properties of differentiated CNVs using high density SNP array among European taurine, African taurine, and indicine groups, and provided a list of lineage-differentiated CNVs, which were involved in traits related to parasite resistance, immunity response, body size, fertility, and milk production<sup>40</sup>. Bickhart *et al.* further explored the diversity and population genetics of both taurine and indicine cattle based on CNV using next generation sequencing and showed hundreds of copy number variable genes were breed-specific<sup>31</sup>. Although a few studies have been carried out to investigate CNV in Chinese cattle<sup>41–43</sup>, genome-wide CNV landscapes and its population-genetic properties in Chinese cattle adapted for local specific environments are largely unknown.

In this study, we performed a genome-wide CNV analysis using high density SNP array in diverse Chinese cattle populations. The objectives of this study were to 1) Generate a comprehensive CNV landscape in Chinese cattle populations; 2) Investigate and compare the diversity and population-genetic properties of CNVs; 3) Explore the diverse selection patterns involved with CNV genes for local adaptation in Chinese native cattle.

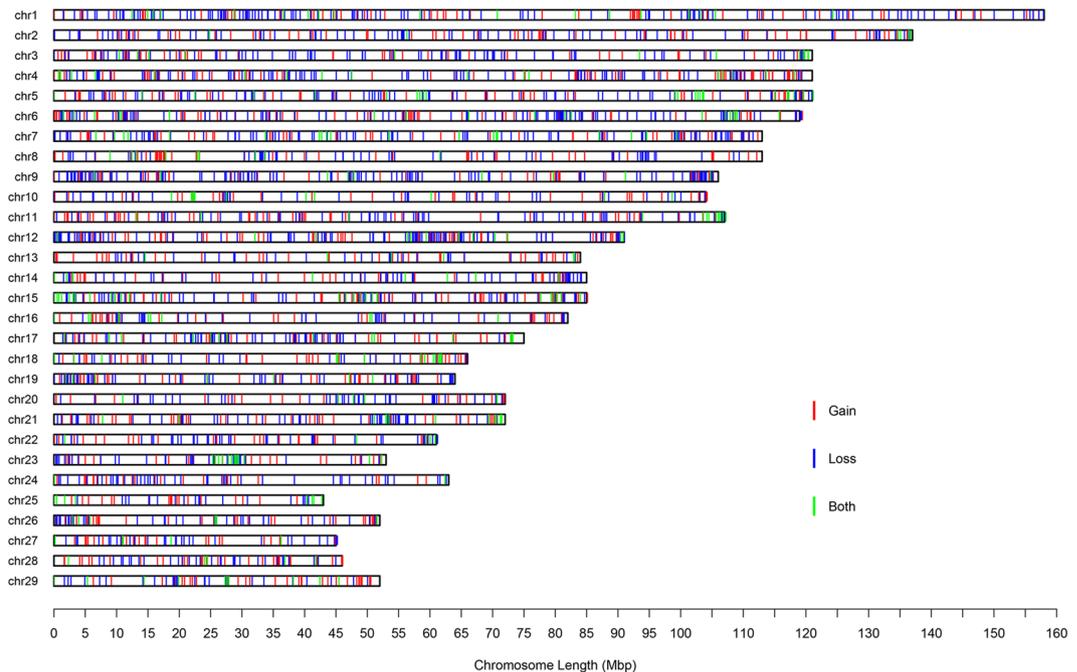
## Results

**CNVs identification.** We performed a genome-wide CNV analysis using the Illumina Bovine HD SNP array in 188 individuals from eight Chinese cattle populations (Fig. 1). After filtering by call rate and genetic relationship, a total of 167 individuals remained. Autosomal CNVs were identified following standard analysis procedures using PennCNV based on the taurine reference assembly (UMD3.1). After CNV detection, low-quality samples were filtered out. A total of 157 individuals were finally used for downstream analyses. Our study detected a total of 13,325 CNVs, representing an average length of 61.2 kb across all individuals (Table 1). These CNVs were merged into 3,356 copy number variant regions (CNVRs), covering 148.0 Mb (average of 44.1 kb) of polymorphic sequence, and corresponding to 5.81% of autosomal genome (148/2,545.9) and 5.07% (148/2,918.0) of the whole genome (Fig. 2 and Supplementary File 1: Table S1). Among them, 2,124 “unique” (only present in one individual), 1,278 “gain” (account for 38.1%), 1,748 “loss” (52.1%) and 330 “both” (9.8%) were identified in our analysis. Notably, we found the count of loss event was approximately 1.4-fold more than gain events, and 5.3-fold more than both events, however, the average length of “both” CNVRs (176.1 kb) were larger than “loss” (32.3 kb) and “gain” (26.2 kb).

Next, we divided the 157 individuals from eight populations into 4 groups which consisted of North (n = 35), Northwest (n = 17), Southwest (n = 63) and South (n = 42) based on the Multidimensional scaling (MDS) analysis results based on SNPs (Figure S1). The CNV events among individuals within each group were merged into

Group	CNV				Sample Size	CNVR					
	CNV	Gain	Loss	Length		CNVR	Unique	Gain	Loss	Both	Length
Total	13,325 (84.87)	4,534	8,791	814,865,357 (61,153)	157	3,356	2,124	1,278	1,748	330	148,099,076 (44,130)
North	2,600 (74.29)	940	1,660	184,528,642 (70,973)	35	969	658	329	557	83	60,838,371 (62,785)
Northwest	1,254 (73.76)	511	743	72,509,715 (57,823)	17	544	373	258	240	46	27,810,707 (51,123)
Southwest	5,472 (86.86)	2,100	3,372	308,854,760 (56,443)	63	1,880	1,259	841	886	153	84,525,087 (44,960)
South	3,999 (95.21)	983	3,016	248,972,240 (62,259)	42	1,401	943	364	934	103	62,497,649 (44,609)

**Table 1.** CNV events and CNVR detected in four groups (North, Northwest, Southwest and South groups). Numbers in parentheses represent per sample averages except in the case of the “Length” column which is the average length. The number of CNVR (Gain, Loss and Both) represent nonredundant CNVR counts.

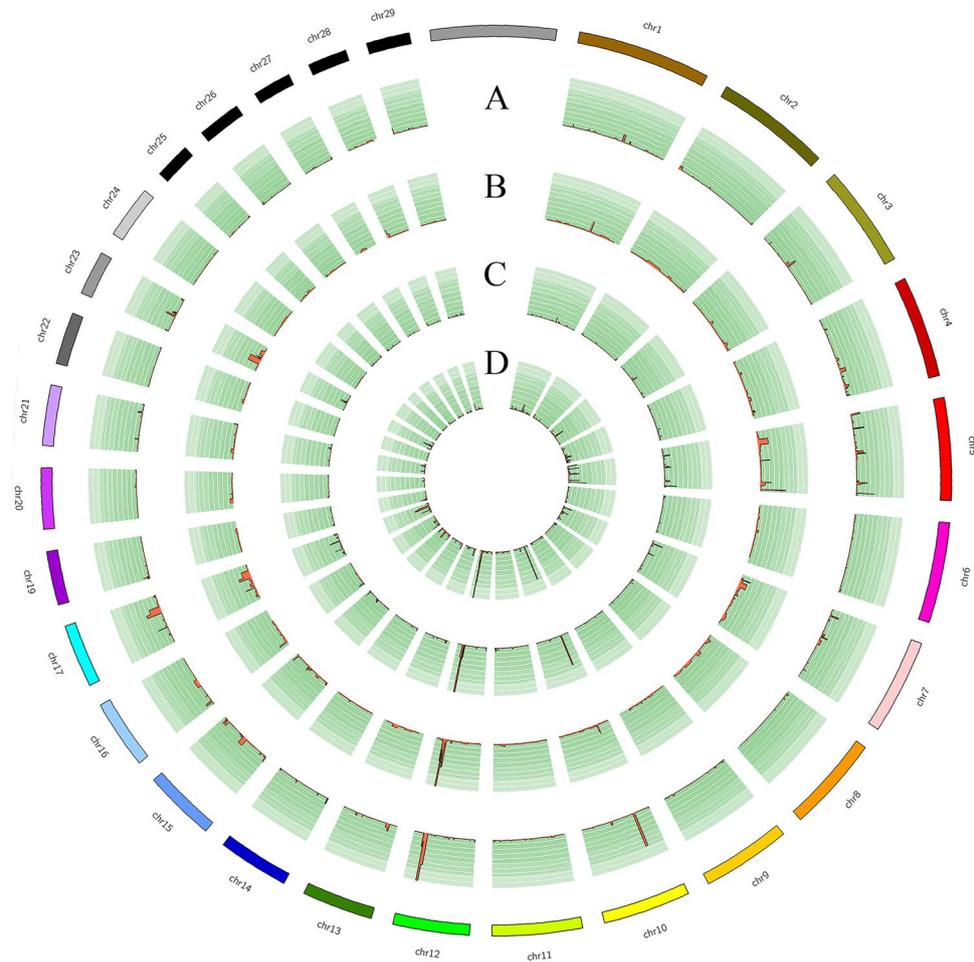


**Figure 2.** Genomic distribution and status of detected CNVRs in 157 cattle. Red, green, and blue lines represent the predicted statuses of gain, loss, and both, respectively. A total of 13,325 CNVs were merged into 3,356 copy number variant regions (CNVRs), covering 148.0 Mb (average of 44.1 kb) of polymorphic sequence. Among them, 2,124 “unique” (only present in one individual), 1,278 “gain” (account for 38.1%), 1,748 “loss” (52.1%) and 330 “both” (9.8%) were identified in our analysis.

group-specific CNVRs. We observed 969, 544, 1880 and 1401 CNVRs in North, Northwest, Southwest and South groups with 62.8, 51.1, 45.0 and 44.6 kb of average length, respectively (Supplementary File 1: Table S1). To display the distribution of CNVR across genome, we filtered away CNVRs with only one CNV event, these single CNV events are more likely to be false-positive CNVs which are not real, and finally 1,232 CNVRs remained. Based on the 1,232 CNVRs, we generated a Circos plot to visualize CNVR landscapes across groups (Fig. 3).

**CNVR annotation and enrichment analysis.** To explore the potential function involved with CNV genes in Chinese cattle, we performed functional annotation and GO enrichment analysis of CNV genes using PANTHER. A total of 1,400 unique genes were found overlapped with 3,356 CNVR, and these identified genes were further used for PANTHER enrichment analysis. We found the identified CNV regions were mainly enriched in fatty acid beta-oxidation, catabolic process, catabolic process, transport, localization, system process, neurological system process and G-protein coupled receptor signaling pathway sensory perception (Supplementary File 2: Table S2). Simultaneously, we carried out the gene feature analysis using the detected CNV genes, and we observed a total of 891 CNVRs and 679 CNVRs overlapping with 1,320 CDSs (coding sequences) and 1,189 exons, respectively.

**Selection signatures of CNVs.** To investigate the group-specific selection involved with CNV events,  $V_i$ , a new statistic modified from  $d_i$ , was proposed to characterize group differentiation of CNVs. The  $V_i$  value was defined as a function of unbiased estimates of all pairwise  $V_{ST}$  between one group and the remaining groups, and  $V_i$  statistic was suitable for detecting selection specific involved with CNVs to a particular group. In this study, we



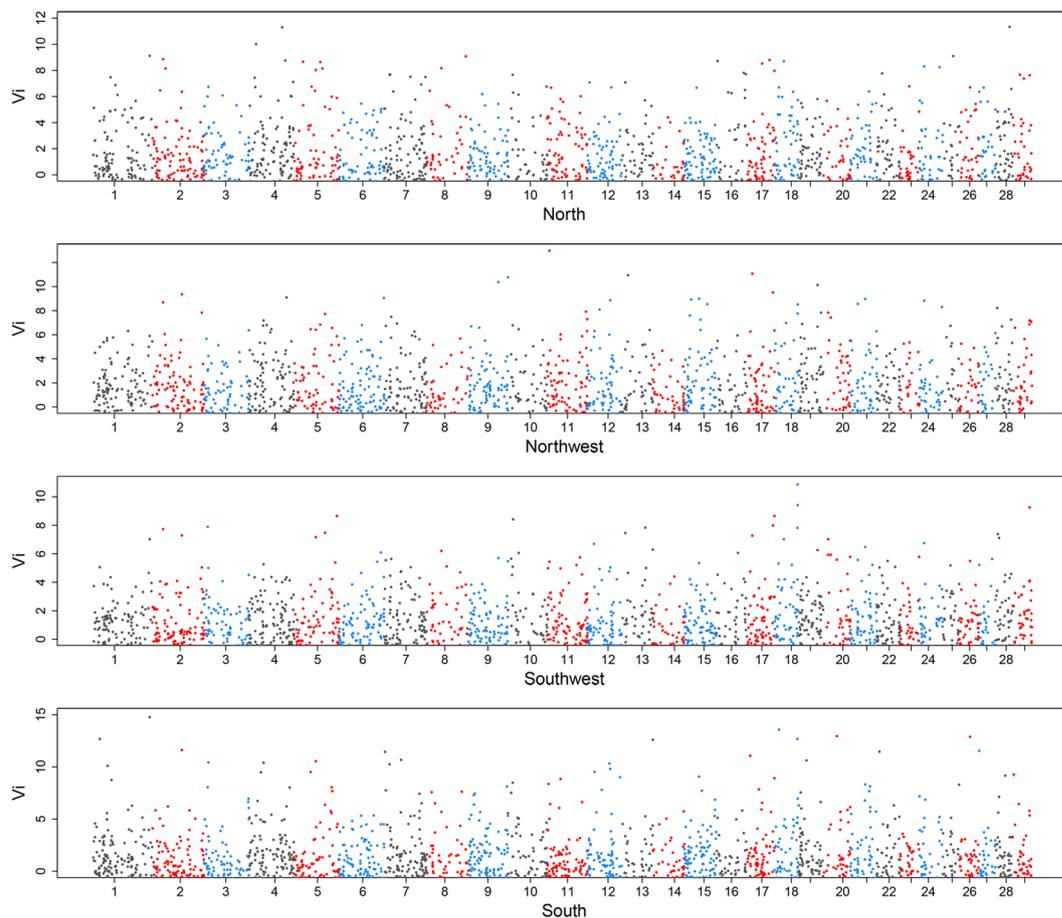
**Figure 3.** Circos plot illustrating CNV regions in 4 groups of cattle. The circles from outside to inside represent frequency of CNV event in each CNVR from North (A), Northwest (B), South (C) and Southwest (D) group respectively.

performed a genome-wide scan for differentiation analysis using  $V_i$  in four groups (North, Northwest, Southwest and South). We then generated Manhattan plots of  $V_i$  distribution for four groups as shown in Fig. 4 (Summary statistics were presented in Supplementary File 3: Table S3). After filtering away CNVRs with only one CNV event, 1,232 CNVRs remained. For each group, we defined candidate regions involved with selection using two thresholds: top 1% or 5% of CNVRs with highest  $V_i$  values in the empirical distribution. In this study, we finally obtained 12 and 62 candidate CNVRs at the top 1% and top 5% for each group, and threshold of  $V_i$  value for top 1% were 6.70, 7.19, 6.07 and 9.25 in North, Northwest, Southwest and South group, while the threshold for top 5% were 3.80, 4.62, 3.32 and 4.95, respectively.

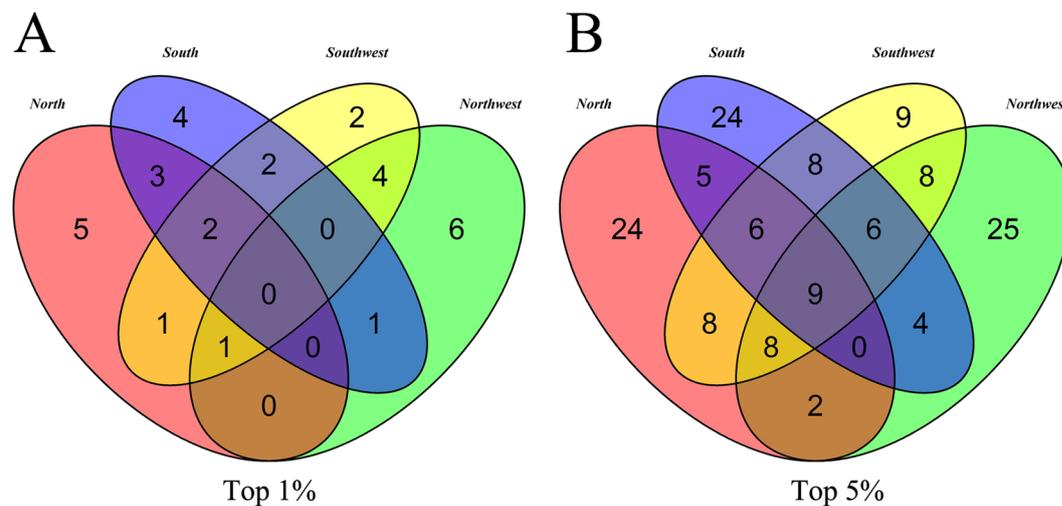
To identify shared and group-specific CNVRs under potential selection, we next generated the Venn diagram based on these CNVRs (Fig. 5). At top 1%, 5, 4, 2 and 6 of group-specific CNVRs were found in North, South, Northwest and Southwest groups, respectively, while no shared CNVR was found. Next, we relaxed the threshold from top 1% to top 5%, we observed 9 CNVRs were shared by all 4 groups, 82 CNVRs were identified as group-specific CNVRs, while 24 CNVRs were detected in North, 24 in South, 25 in Northwest and 9 in Southwest group, respectively. To further estimate the divergence of CNVs across groups for the identified candidate CNVRs, we extracted the LRR values for each shared and unique CNVRs in four groups. Notably, we observed clear differences based on average LRR for each group-specific CNVR and the Box-plot of the distributions of average LRR in the identified CNVRs across diverse cattle groups were shown in Fig. 6.

**Genes in shared and group-specific CNVRs.** To investigate lineage-differentiated CNV genes in Chinese cattle, we also conducted CNVR-based population differentiation analysis, and identified a set of potential CNV genes under divergent selection (Table 2). In the shared CNVRs, we detected 4 genes embedded with 3 CNVRs, among them, CNVR208 and CNVR2872 on BTA1 and BTA22 with length of 14.6 and 15.0 kb, overlapping gene *CBR1* and *TKT*, respectively, and CNVR2451 on BTA17 with length of 71.4 kb spanning two genes *LOC100297192* and *IGLL1*.

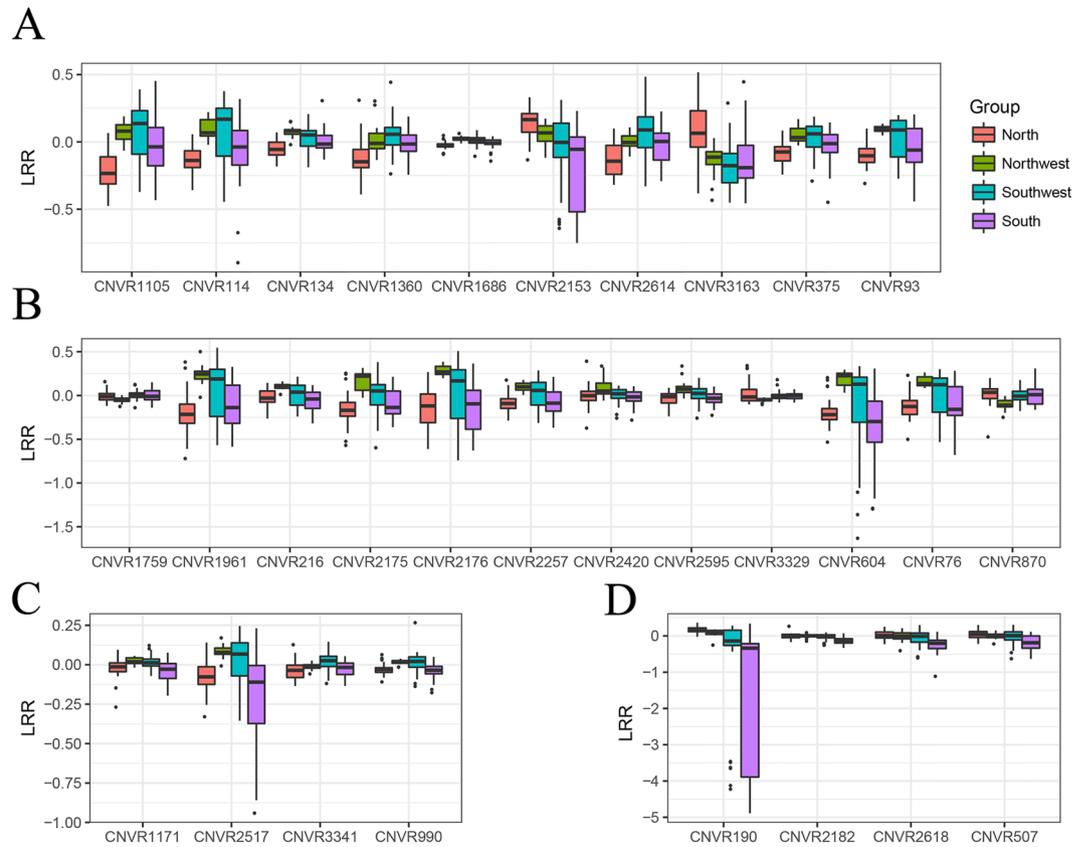
In this study, a total of 30 group-specific CNVRs were overlapped with 46 annotated genes for 4 groups (Table 2). In North group, we obtained 10 CNVRs covering 15 genes, which were mainly associated with metabolic process (*GUCY1A2*, *RAVER1*, *EBD*, *DEFB1* and *DEFB5*) and response to stimulus (*GUCY1A2*, *EBD*,



**Figure 4.** Genome-wide distribution of  $V_i$  values among North, Northwest, Southwest and South groups. The  $V_i$  value was defined as a function of unbiased estimates of all pairwise  $V_{ST}$  between one group and the remaining groups within a population, and  $V_i$  statistic was suitable for detecting selection specific involved with CNVs to a particular group. The distribution of  $V_i$  for each CNVR across all auto chromosomes is shown for each group. Alternating color indicates  $V_i$  values from adjacent chromosomes.



**Figure 5.** Discovery of shared and group-specific CNVRs among 4 groups. The Venn diagram shows the number of overlapping CNVRs in 4 groups including North, Northwest, Southwest and South groups. The top 1% ( $n = 12$ ) and top 5% ( $n = 60$ ) of CNVRs among four groups were displayed in left and right, respectively.



**Figure 6.** Box-plot of average LRR values for group-specific CNVRs across four groups. (A) North-specific group, (B) Northwest-specific group, (C) Southwest-specific group, (D) South-specific group.

*DEFB1*, *DEFB5* and *ABCC5*). Moreover, gene *TMC6* was related to immune system in CNVR2614 ( $V_i = 4.42$ ), and *COL27A1* within CNVR1360 ( $V_i = 4.44$ ) and *BOC* gene within CNVR114 ( $V_i = 4.97$ ) were identified as North-specific candidates. In addition, we detected several genes including calcium ion binding (*EFHB*), phosphatase activity (*DUSP27*) and GTP binding and phosphorus-oxygen lyase activity (*GUCY1A2*). In Northwest group, a total of 15 genes were obtained in 12 CNVRs, among them, most genes were related to metabolic process (*KIN*, *RPL7A*, *SATB1*, *ABO* and *FMOD*), cellular process (*KIN*, *GLRB*, *SATB1* and *FMOD*), response to stimulus (*KIN* and *GLRB*) and developmental process (*EPHA3* and *FMOD*). Gene *MRV11* was simultaneously overlapped with CNVR2175 and CNVR2176 with  $V_i = 6.40$  and  $V_i = 7.25$ . For Southwest group, we detected 8 genes in 4 group-specific CNVRs. The genes *FAT2* and *SLC36A1* overlapped with CNVR1171 ( $V_i = 3.78$ ), and CNVR3341 overlapped with three genes including *LRFN4*, *MIR2408*, *PC* and *RCE1*. We also obtained CNVR990 and CNVR2517, which overlapped with gene *KIT* and *ZNF331*, respectively. *KIT* was identified in CNVR990 ( $V_i = 3.40$ ) in Southwest group, which was also shown to have high  $V_i$  value (4.49) in South group. We found 4 group-specific CNVRs with 8 genes in South group, most of these genes involved with cellular process, biological regulation and response to stimulus. We observed CNVR2181 with four genes showing relatively high value ( $V_i = 7.71$ ), CNVR190 ( $V_i = 5.14$ ) and CNVR2618 ( $V_i = 6.63$ ) overlapped with genes *KPNA6* and *ATP5H*, while CNVR507 overlapped with two genes (*ESPNL* and *SCLY*) (Table 2).

## Discussion

In this study, we performed a genome-wide CNV scan using high density SNP array in Chinese cattle. Recently, many studies regarding CNV discovery had been reported for various cattle populations using aCGH, SNP array and next generation sequencing<sup>25,28,31,39,44–48</sup>. For instance, some previous studies carried out CNV analysis in world-wide cattle including taurine, indicine, and mixed cattle population from African using both BovineSNP50 and BovineHD SNP array<sup>26,44</sup>. In addition, other studies have conducted CNV analyses for local cattle population, including South African Nguni cattle<sup>49</sup>, Brazil Nelore cattle<sup>50,51</sup>, Hanwoo and Japanese Black cattle<sup>52,53</sup>. For Chinese native cattle, only a few studies have been carried out using aCGH and low density SNP array in limited populations<sup>41–43</sup>. Therefore, a comprehensive investigation of CNV and their population genetic properties in diverse Chinese cattle is needed. In the current study, CNV analysis for CDM, WSC, ZTC and NDC was explored for the first time, and the newly discovered CNVs in specific local populations could offer important molecular resources and may further help to elucidate the selection mechanisms of CNV and their genomic adaptation.

In the current study, we detected 13,225 CNV events and 3,356 CNVRs across the cattle genome, of which, 90.33 Mb were newly detected as compared to the cattle HapMap samples<sup>31,44</sup>. In contrast to previous CNV studies in Chinese cattle, we observed 11.38 Mb CNVR was overlapped with previous reports and 136.72 Mb

Group	CNVR	Chr	Start	End	$V_i$ -North	$V_i$ -Northwest	$V_i$ -Southwest	$V_i$ -South	Candidate Genes
Shared	208	1	150050992	150065547	9.11	5.90	7.03	14.76	CBR1
	2451	17	72870233	72941663	7.96	6.36	8.65	8.91	IGLL1,LOC100297192
	2872	22	48265166	48280157	4.22	4.78	5.07	5.02	TKT
North	93	1	44683130	44743553	4.89	4.02	1.67	0.74	EFHB
	114	1	58435550	58449521	4.97	2.37	2.20	0.44	BOC
	134	1	83760081	83775860	4.99	3.34	1.77	1.00	ABCC5
	375	3	1667959	1704027	5.24	1.80	2.44	0.73	DUSP27
	1105	7	16054798	16073843	7.66	1.65	2.65	1.12	FDX1L,ICAM5,RAVER1
	1360	8	105256615	105273649	4.44	-0.95	3.10	0.82	COL27A1
	1686	11	38648841	38851335	4.32	2.85	3.07	1.60	CCDC85A
	2153	15	17038087	17074274	4.00	0.51	1.85	4.88	GUCY1A2
	2614	19	54639709	54648592	4.42	-0.06	2.58	0.52	TMC6
3163	27	5444267	5973102	5.60	-0.88	1.76	0.01	DEFB,DEFB1,DEFB5,EBD	
Northwest	76	1	36710870	36729816	2.72	5.52	1.56	1.51	EPHA3
	216	1	157196044	157225181	0.53	5.44	1.17	2.31	SATB1
	604	4	46975403	46982930	2.77	6.81	0.97	2.12	ATXN7L1
	870	5	119430708	119531892	-0.77	5.15	2.13	-0.95	FAM19A5
	1759	11	104195124	104384812	-1.18	4.90	2.74	-1.26	ABO,MED22,RPL7A,SURF4,SURF6
	1961	13	16187783	16193492	3.24	5.43	2.73	2.86	KIN
	2175	15	42588302	42594246	3.08	6.40	2.71	2.71	MRV11
	2176	15	42639417	42644258	2.65	7.25	3.14	3.29	MRV11
	2257	16	933282	950232	2.28	4.89	1.88	1.96	FMOD
	2420	17	43069432	43085729	-1.14	5.01	2.17	1.08	GLRB
	2595	19	41945417	41951980	0.89	4.86	2.42	2.41	LOC618938
	3329	29	37219773	37404822	-0.72	4.88	2.00	0.41	MIR584-7
	Southwest	990	6	71630923	7207236	2.09	2.75	3.40	4.49
1171		7	64638445	64705407	0.17	0.78	3.78	4.51	FAT2,SLC36A1
2517		18	60990841	61031513	3.12	2.86	4.19	4.62	ZNF331
3341		29	45491233	45608617	0.57	-0.68	4.11	0.56	LRFN4,MIR2408,PC,RCE1
South	190	1	130623810	130626113	3.11	1.21	2.26	5.14	KPNA6
	507	3	118059058	118090512	1.51	-0.50	1.52	6.66	ESPNL,SCLY
	2182	15	46571384	47009702	1.33	0.08	2.31	7.71	LOC100125776,LOC506989,LOC530354,MRPL17
	2618	19	56993243	57020892	1.97	-0.68	1.35	6.63	ATP5H

**Table 2.** Summary statistics of shared and group-specific CNVRs, including the position of CNVR, estimated  $V_i$  value for each group and corresponding CNV genes. Candidate genes were defined as the genes overlapping with CNVR.

CNVR was newly discovered in present study<sup>41-43</sup>. In addition, we also found 54.76 Mb CNVRs were overlapped as compared to the recent publication in Nelore cattle and 93.34 Mb CNVR were specifically detected in current study<sup>54</sup>. Totally, we detected 4,534 gain and 8,791 loss events across cattle populations in China, and the results shown that deletions are more numerous than duplications which is generally observed in human<sup>55</sup>, mouse<sup>11</sup>, dog<sup>15,56</sup>, and cattle<sup>25,44,45</sup>. This finding may indicate that deletions were more easily to be detected by PennCNV on the BovineHD SNP array<sup>44</sup>. We observed more CNV events were detected in South and Southwest (95.21 and 86.86 per sample) than in North (74.29 per sample) and Northwest (73.76 per sample), and a previous study also detected more CNV events in indicine breed than taurine<sup>26</sup>. This may be because Northern group was more influenced by *B. taurus*, and South group was more influenced by *B. indicus*, while the cattle in the central and lower areas of the Yellow River and the Huaihe River appear to be a mixture of *B. taurus* and *B. indicus* types<sup>57,58</sup>. This finding may imply subspecies divergence of genome structures in Chinese cattle. In addition, some of these differences could be related to the fact that SNP markers in BovineHD were designed based on the UMD3.1 reference genome, which may cause bias for CNV calling in indicine-derived cattle populations<sup>45</sup>.

Diversity and selection aspects of CNVs have been extensively explored in world-wide cattle populations using aCGH, SNP array and next generation sequencing, demonstrating that CNVs display breed-specific differences and may be associated with adaptation, health, and production traits<sup>25,31,39,40,46</sup>. To our knowledge, our study is a first attempt to explore the CNV properties using high density SNP array in Chinese cattle across a broad latitudinal range. We suspect CNV are important genomic variants under distinct selection pressures, and CNV can contribute to diverse morphology in cattle sampled from different geographic regions with various environmental factors (e.g., climate, temperature, altitude, rainfall, and food source). Our study explored lineage-differentiated CNVs in Chinese cattle and identified several potential CNV candidates under divergent selection for adaptation in local environmental conditions. CNV genes may have distinct functional roles and be subject to different evolutionary

pressures. Interestingly, for shared CNVR with high  $V_i$  values across four groups (North, Northwest, Southwest and South), we identified several functionally important genes related to immunity and metabolism, such as *IGLL1*, *CBRI*, and *TKT*. *IGLL1* has been previously detected with copy number changes in cattle genome<sup>40,44</sup>, this gene has been showed differentially expressed in hepatic and mammary tissue in dairy cows<sup>59,60</sup> and was associated with resistance to gastrointestinal nematodes in Angus cattle<sup>61</sup>. The expression of *CBRI* was found to be associated with oxidative stress in bovine embryos<sup>62</sup>, and this gene had been reported to be associated with growth traits weaning gain in *Bos indicus*<sup>51</sup>. Previous studies revealed that *TKT* involved in fatty acid synthesis and storage in muscle, which were also correlated with intramuscular fat in both cattle and sheep<sup>63</sup>. The shared CNVs across cattle groups could suggest the parallel selection on CNVs through evolution of cattle genomes, which was also reported in three-spined stickleback<sup>64</sup>. The existence of shared CNVs indicates their potential function may contribute to the parallel adaptive evolution within multiple natural populations in diverse species.

Moreover, our study revealed several group-specific CNV genes, which may play significant roles in diverse morphology and adaptation for local environmental condition. In the North group, we identified several candidate genes related to calcification of cartilage and immune system in MGC and YHC. *TMC6* related to immune system in CNVR2614 ( $V_i = 4.42$ ) was associated with milk somatic cell score in dairy cattle, which may imply the potential selection for this gene related to the milk quality in cold environmental condition<sup>65</sup>. *COL27A1* within CNVR1360 ( $V_i = 4.44$ ) may play a crucial role in cartilage calcification and average daily gain<sup>66,67</sup>. Our results indicate these CNV genes could enhance development of bone and muscle, cold tolerance and disease resistance for North group cattle, which live in relatively cold environment in northern China.

For the Northwest group, CDM cattle live on the Qinghai-Tibet plateau in northwestern China, which is an arid environment exhibiting dry, hypoxia, low air pressure. Specifically, we observed several genes *MRVII*, *ABO*, *GLRB* and *EPHA3* related to nervous system, platelet reactivity, parasite resistance and histoblood group antigens. *MRVII* overlapped with both CNVR2175 and CNVR2176, and several variants in *MRVII* had been reported to be associated with platelet count, mean platelet volume and platelet reactivity<sup>68</sup>. In the current study, *ABO*, embedded within CNVR1759 ( $V_i = 4.90$ ), is related to A and B histoblood group antigens<sup>69</sup>, and copy number changes have been identified within *ABO* in recent studies by Hou *et al.*<sup>26,61</sup>. *ABO* have also been associated with parasite resistance and susceptibility to gastrointestinal nematodes in Angus cattle<sup>61</sup>. Moreover, a previous study suggested one CNVR containing *ABO*, *SURF6*, *RPL7A* was significantly associated with milk somatic cell score in Holstein cattle<sup>70</sup>. *GLRB*, which was identified as candidate in CDM cattle at BTA17 has been associated with flight speed in beef steers, and *GLRB* protein is a ligand gated ion-channel subunit throughout the central nervous system<sup>71,72</sup>.

In the Southwest group (PWC, LSC and ZTC), we identified several genes *KIT*, *FAT2*, *SLC36A1*, *ZNF331*, *LRFN4* and *RCE1* in four south-specific CNVRs. Notably, *KIT* is a tyrosine kinase receptor, and normal *KIT* signaling is required for development and survival of neural crest-derived melanoblasts<sup>73</sup>. Genetic variation in *KIT* gene has been shown to affect coat coloring pattern in a variety of mammals<sup>74-77</sup>. Rubin *et al.* suggested that white and white spotted pigs are caused by at least two out of the four *KIT* duplications and the porcine *KIT* locus also illustrates the evolution of alleles under strong positive selection<sup>78</sup>. Our results suggested the identified genes may facilitate the cattle living in mountainous areas of Southwestern China.

This study provides a comprehensive investigation of CNV properties in diverse Chinese cattle, and newly identified CNVs contribute to the important genetic resources for the global cattle population. Our results further suggest that lineage-differentiated CNVs may be under divergent selection for adaptation in local environmental conditions. Hybridization array studies may generate both false positive and false negative results, regardless of how the data are analyzed for CNV discovery<sup>79</sup>. Many studies recommend using multiple CNV calling algorithms instead of just one<sup>80</sup>; however, although the net effect of this strategy decreases the false negative rate, it also increases the false positive rate<sup>81</sup>. With advances in next generation sequencing projects, such as the 1000 Human Genomes project<sup>2</sup> and the 1000 Bull Genomes project<sup>29</sup>, we should be able to better estimate the false positive and false negative rates with better CNV calling standards. Therefore, careful experimental design and rigorous data filtering were required to reveal the impacts of CNVs on both phenotypic variability and diverse selection. Future CNV studies utilizing next-generation sequencing (NGS) and complementary analysis programs will help precisely define the CNV structure and elucidate its function.

## Materials and Methods

**Ethics Statement.** All of the animal experiments were approved by the Chinese Academy of Agricultural Sciences (CAAS, Beijing, China). All of the animal procedures were performed in strict accordance with the guidelines proposed by the China Council on Animal Care and the Ministry of Agriculture of People's Republic of China.

**Sample selection.** Blood samples were obtained from 188 individuals representing 8 different cattle breeds across a broad latitudinal range in China. These include Menggu cattle (MGC), Yanhuang cattle (YHC), Caidamu cattle (CDM), Pingwu cattle (PWC), Liangshan cattle (LSC), Zhaotong cattle (ZTC), Wenshan cattle (WSC), and Nandan cattle (NDC). Genomic DNA was extracted from blood samples using the TIANamp Blood DNA Kit (Tiangen Biotech Co. Ltd), and DNA with an A260/280 ratio ranging between 1.8 and 2.0 was subject to further analysis. In this study, we divided 8 breeds into 4 groups based on geographical locations (North group, Northwest group, Southwest group and South group) (Fig. 1). The genotyping platform adopted in this study was Illumina's Infinium II Multi-Sample Assay. SNP chips were scanned using iScan and analyzed using Illumina's GenomeStudio 2011. After filtering by the call rate of each given animal (threshold was > 95%), the final data including Log R Ratio (LRR) and B Allele Frequency (BAF) were exported from GenomeStudio software. To avoid the bias of population genetic estimation, genetic relationships between pairwise individuals were estimated using PI-HAT value implemented in PLINK v1.07<sup>82</sup>, unrelated individuals with pairwise PI-HAT < 0.25 were kept for subsequent analyses.

**CNVs detection.** In this study, we utilized the PennCNVv1.0.3 software to detect CNV across autosomes in Chinese cattle populations<sup>83</sup>. chrX and chrUn were not considered here due to mapping uncertainty as described previously by Hou *et al.*<sup>26</sup>. The PennCNV algorithm incorporates LRR and BAF, which denote the normalized intensity ratio for each SNP alleles and the frequency of allele B, respectively. The population frequency of B allele (PFB) file was calculated based on the BAF of each marker across populations. The gcmodel file was generated by calculating the GC content of the 1 Mbp genomic region surrounding each marker (500 kb each side). The final CNVs were obtained by filtering the low quality samples with the following thresholds: standard deviation (SD) of LRR as 0.35, BAF drift as 0.01 and waviness factors as 0.05.

**CNVR compilation.** CNV regions (CNVRs) were produced by aggregating overlapping CNVs (by at least 1bp) across samples using BEDTools v2.26.0<sup>84</sup>. The CNVRs were classified as “gain”, “loss”, or “both” events. Overlapping “loss” and “gain” CNVRs were merged into single regions to account for “both” events. To facilitate the comparison of CNV pattern among the diverse groups, the CNVRs for each group were generated. The frequency landscape of each group-specific CNVR was visualized using Circos software<sup>85</sup>.

**Function annotation.** To elucidate the functional aspect involved with identified CNVs in cattle genome, gene content of cattle CNV regions was assessed using the RefGene track of the UCSC genome browser. Annotation of genes and gene feature analysis (identify the coding sequence and exon of genes presented within CNVs) was performed using the scan\_region.pl script from PennCNV v1.0.3 package. PANTHER (Protein ANalysis THrough Evolutionary Relationships) classification system (<http://www.pantherdb.org/>) was used to explore functional ontology categories as described previously<sup>86</sup>. We tested the hypothesis that the PANTHER molecular function, biological process and pathway terms were under- or over- represented in CNV regions after Bonferroni corrections. Only GO terms with P-value < 0.05 after the Bonferroni correction were considered.

**Signatures of selection.** To detect the group-specific CNV events, we proposed a statistic named  $V_i$  to estimate the region-specific divergence in CNVR for each group based on unbiased estimates of pairwise  $V_{ST}$ , which estimates population differentiation based on average LRR values across all probes falling within a specific CNV region<sup>87</sup>.

For each CNVR, we calculated the statistic  $V_i = \frac{V_{st}^{ij} - E[V_{st}^{ij}]}{sd[V_{st}^{ij}]}$ , where  $E[V_{st}^{ij}]$  and  $sd[V_{st}^{ij}]$  denote the expected value and standard deviation of  $V_{ST}$  between groups  $i$  and  $j$  calculated from the average LRR values across all CNVRs<sup>87</sup>.  $V_{ST}$  is calculated using the following equation:  $(V_T - V_S)/V_T$ , where  $V_T$  is the variance in LRR apparent among all unrelated individuals and  $V_S$  is the average variance within each group, weighted for sample size. The  $V_i$  was inspired by statistic  $d_i$  proposed by Akey *et al.*<sup>88</sup>, which is based on the estimation of difference of allele frequency across group.  $d_i$  measures the standardized locus-specific deviation using SNP genotype in levels of population structure for a particular group relative to the genome-wide average. Here, we proposed  $V_i$  to measure the changes of copy number variant regions (based on average LRR values) for a particular group. Large positive  $V_i$  values indicate high levels of group difference of CNV regions relative to the genome-at-large, which are potentially involved with selection. Thus,  $V_i$  is particularly well suited for detecting selection of CNV region specific to a particular group.

The Manhattan plot of  $V_i$  value for each CNVR across groups were created by the R package “qqman”. To identify the shared and group-specific CNVRs, the Venn diagram was generated based on overlapping of CNVRs with R package “VennDiagram”<sup>89</sup>. Box-plot was generated to display the distribution of average LRR in the identified CNVRs across diverse cattle groups. Unless specified, all statistical analyses were performed using R programming (<https://www.R-project.org>).

## References

- Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature genetics* **39**, S7–15, <https://doi.org/10.1038/ng2093> (2007).
- Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65, <https://doi.org/10.1038/nature09708> (2011).
- Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81, <https://doi.org/10.1038/nature15394> (2015).
- Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646, <https://doi.org/10.1126/science.1197005> (2010).
- Zhang, F., Gu, W., Hurler, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics* **10**, 451 (2009).
- Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nature reviews. Genetics* **16**, 172–183, <https://doi.org/10.1038/nrg3871> (2015).
- Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature reviews. Genetics* **10**, 551–564, <https://doi.org/10.1038/nrg2593> (2009).
- Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712, <https://doi.org/10.1038/nature08516> (2010).
- Moreira, E. F., Larrayoz, I. M., Lee, J. W. & Rodriguez, I. R. 7-Ketocholesterol is present in lipid deposits in the primate retina: potential implication in the induction of VEGF and CNV formation. *Investigative ophthalmology & visual science* **50**, 523–532, <https://doi.org/10.1167/iovs.08-2373> (2009).
- Cutler, G. & Kassner, P. D. Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. *Cytogenetic and genome research* **123**, 297–306 (2008).
- Graubert, T. A. *et al.* A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS genetics* **3** (2007).
- Cutler, G., Marshall, L. A., Chin, N., Baribault, H. & Kassner, P. D. Significant gene content variation characterizes the genomes of inbred mouse strains. *Genome research* **17**, 1743–1754 (2007).
- Brown, K. H. *et al.* Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 529–534, <https://doi.org/10.1073/pnas.1112163109> (2012).

14. Chen, W.-K., Swartz, J. D., Rush, L. J. & Alvarez, C. E. Mapping DNA structural variation in dogs. *Genome research* **19**, 500–509 (2009).
15. Nicholas, T. J. *et al.* The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome research* **19**, 491–499 (2009).
16. Nicholas, T. J., Baker, C., Eichler, E. E. & Akey, J. M. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC genomics* **12**, 1 (2011).
17. Wang, X., Nahashon, S., Feaster, T. K., Bohannon-Stewart, A. & Adefope, N. An initial map of chromosomal segmental copy number variations in the chicken. *BMC genomics* **11**, 1 (2010).
18. Völker, M. *et al.* Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome research* **20**, 503–511 (2010).
19. Fadista, J., Nygaard, M., Holm, L.-E., Thomsen, B. & Bendixen, C. A snapshot of CNVs in the pig genome. *PLoS one* **3**, e3916 (2008).
20. Ramayo-Caldas, Y. *et al.* Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC genomics* **11**, 1 (2010).
21. Zhu, C. *et al.* Genome-wide detection of CNVs in Chinese indigenous sheep with different types of tails using ovine high-density 600K SNP arrays. *Scientific reports* **6** (2016).
22. Liu, J. *et al.* Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. *BMC genomics* **14**, 1 (2013).
23. Jenkins, G. M. *et al.* Copy number variants in the sheep genome detected using multiple approaches. *BMC genomics* **17** (2016).
24. Fontanesi, L. *et al.* An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC genomics* **11**, 1 (2010).
25. Liu, G. E. *et al.* Analysis of copy number variations among diverse cattle breeds. *Genome research* **20**, 693–703 (2010).
26. Hou, Y. *et al.* Genomic characteristics of cattle copy number variations. *BMC genomics* **12**, 127, <https://doi.org/10.1186/1471-2164-12-127> (2011).
27. Bae, J. S. *et al.* Identification of copy number variations and common deletion polymorphisms in cattle. *BMC genomics* **11**, 232, <https://doi.org/10.1186/1471-2164-11-232> (2010).
28. Zhan, B. *et al.* Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC genomics* **12**, 557, <https://doi.org/10.1186/1471-2164-12-557> (2011).
29. Bickhart, D. M. *et al.* Copy number variation of individual cattle genomes using next-generation sequencing. *Genome research* **22**, 778–790 (2012).
30. Iskow, R. C., Gokcumen, O. & Lee, C. Exploring the role of copy number variants in human adaptation. *Trends in genetics: TIG* **28**, 245–257, <https://doi.org/10.1016/j.tig.2012.03.002> (2012).
31. Bickhart, D. M. *et al.* Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. *DNA research: an international journal for rapid publication of reports on genes and genomes* **23**, 253–262, <https://doi.org/10.1093/dnares/dsw013> (2016).
32. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761, <https://doi.org/10.1126/science.aab3761> (2015).
33. Chain, F. J. *et al.* Extensive copy-number variation of young genes across stickleback populations. *PLoS genetics* **10**, e1004830, <https://doi.org/10.1371/journal.pgen.1004830> (2014).
34. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature genetics* **39**, 1256–1260, <https://doi.org/10.1038/ng2123> (2007).
35. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440, <https://doi.org/10.1126/science.1101160> (2005).
36. Xue, Y. *et al.* Adaptive evolution of UGT2B17 copy-number variation. *American journal of human genetics* **83**, 337–346, <https://doi.org/10.1016/j.ajhg.2008.08.004> (2008).
37. Waszak, S. M. *et al.* Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS computational biology* **6**, e1000988, <https://doi.org/10.1371/journal.pcbi.1000988> (2010).
38. Hasin, Y. *et al.* High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS genetics* **4**, e1000249, <https://doi.org/10.1371/journal.pgen.1000249> (2008).
39. Keel, B. N., Lindholm-Perry, A. K. & Snelling, W. M. Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome. *Frontiers in genetics* **7**, 207, <https://doi.org/10.3389/fgene.2016.00207> (2016).
40. Xu, L. *et al.* Population-genetic properties of differentiated copy number variations in cattle. *Scientific reports* **6**, 23161, <https://doi.org/10.1038/srep23161> (2016).
41. Zhang, L. *et al.* Detection of copy number variations and their effects in Chinese bulls. *BMC genomics* **15**, 480, <https://doi.org/10.1186/1471-2164-15-480> (2014).
42. Zhang, L. *et al.* Impact of Parental *Bos taurus* and *Bos indicus* Origins on Copy Number Variation in Traditional Chinese Cattle Breeds. *Genome biology and evolution* **7**, 2352–2361, <https://doi.org/10.1093/gbe/evv151> (2015).
43. Zhang, Q., Ma, Y., Wang, X., Zhang, Y. & Zhao, X. Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. *Molecular genetics and genomics: MGG* **290**, 319–327, <https://doi.org/10.1007/s00438-014-0923-4> (2015).
44. Hou, Y. *et al.* Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. *BMC genomics* **13**, 376, <https://doi.org/10.1186/1471-2164-13-376> (2012).
45. Hou, Y. *et al.* Genomic characteristics of cattle copy number variations. *BMC genomics* **12**, 127 (2011).
46. Bickhart, D. M. *et al.* Copy number variation of individual cattle genomes using next-generation sequencing. *Genome research* **22**, 778–790, <https://doi.org/10.1101/gr.133967.111> (2012).
47. Fadista, J., Thomsen, B., Holm, L. E. & Bendixen, C. Copy number variation in the bovine genome. *BMC genomics* **11**, 284, <https://doi.org/10.1186/1471-2164-11-284> (2010).
48. Stothard, P. *et al.* Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC genomics* **12**, 559, <https://doi.org/10.1186/1471-2164-12-559> (2011).
49. Wang, M. D., Dzama, K., Hefer, C. A. & Muchadeyi, F. C. Genomic population structure and prevalence of copy number variations in South African Nguni cattle. *BMC genomics* **16**, 894, <https://doi.org/10.1186/s12864-015-2122-z> (2015).
50. da Silva, J. M. *et al.* Genome-wide copy number variation (CNV) detection in Nelore cattle reveals highly frequent variants in genome regions harboring QTLs affecting production traits. *BMC genomics* **17**, 454, <https://doi.org/10.1186/s12864-016-2752-9> (2016).
51. Zhou, Y. *et al.* Genome-wide CNV analysis reveals variants associated with growth traits in *Bos indicus*. *BMC genomics* **17**, 419, <https://doi.org/10.1186/s12864-016-2461-4> (2016).
52. Sasaki, S., Watanabe, T., Nishimura, S. & Sugimoto, Y. Genome-wide identification of copy number variation using high-density single-nucleotide polymorphism array in Japanese Black cattle. *BMC genetics* **17**, 1–9 (2016).
53. Choi, J. W. *et al.* Copy number variations in Hanwoo and Yanbian cattle genomes using the massively parallel sequencing data. *Gene* **589**, 36–42, <https://doi.org/10.1016/j.gene.2016.05.017> (2016).
54. Silva, V. H. *et al.* Genome-Wide Detection of CNVs and Their Association with Meat Tenderness in Nelore Cattle. *PLoS one* **11**, e0157711, <https://doi.org/10.1371/journal.pone.0157711> (2016).
55. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
56. Berglund, J. *et al.* Novel origins of copy number variation in the dog genome. *Genome biology* **13** (2012).

57. Li, R. *et al.* Paternal origins of Chinese cattle. *Animal genetics* **44**, 446–449, <https://doi.org/10.1111/age.12022> (2013).
58. Cai, X., Chen, H., Wang, S., Xue, K. & Lei, C. Polymorphisms of two Y chromosome microsatellites in Chinese cattle. *Genetics, selection, evolution: GSE* **38**, 525–534, <https://doi.org/10.1051/gse:2006019> (2006).
59. Ouattara, B., Bissonnette, N., Duplessis, M. & Girard, C. L. Supplements of vitamins B9 and B12 affect hepatic and mammary gland gene expression profiles in lactating dairy cows. *BMC genomics* **17**, 640, <https://doi.org/10.1186/s12864-016-2872-2> (2016).
60. Li, C. *et al.* RNA-Seq reveals 10 novel promising candidate genes affecting milk protein concentration in the Chinese Holstein population. *Scientific reports* **6**, 26813, <https://doi.org/10.1038/srep26813> (2016).
61. Hou, Y. *et al.* Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Functional & integrative genomics* **12**, 81–92, <https://doi.org/10.1007/s10142-011-0252-1> (2012).
62. Sakatani, M. *et al.* Changes in the transcriptome of morula-stage bovine embryos caused by heat shock: relationship to developmental acquisition of thermotolerance. *Reproductive biology and endocrinology: RB&E* **11**, 3, <https://doi.org/10.1186/1477-7827-11-3> (2013).
63. Guo, B. *et al.* A gene expression estimator of intramuscular fat percentage for use in both cattle and sheep. *J Anim Sci Biotechnol* **5**, 35, <https://doi.org/10.1186/2049-1891-5-35> (2014).
64. Hirase, S., Ozaki, H. & Iwasaki, W. Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes. *BMC genomics* **15**, 735, <https://doi.org/10.1186/1471-2164-15-735> (2014).
65. Durán Aguilar, M. *et al.* Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. *Journal of animal breeding and genetics = Zeitschrift für Tierzucht und Zuchtungsbiologie*, <https://doi.org/10.1111/jbg.12238> (2016).
66. Pace, J. M., Corrado, M., Missero, C. & Byers, P. H. Identification, characterization and expression analysis of a new fibrillar collagen gene, COL27A1. *Matrix biology: journal of the International Society for Matrix Biology* **22**, 3–14 (2003).
67. Wang, K. *et al.* Genome Wide Association Analysis Reveals New Production Trait Genes in a Male Duroc Population. *PLoS one* **10**, e0139207, <https://doi.org/10.1371/journal.pone.0139207> (2015).
68. Eicher, J. D. *et al.* Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 Individuals. *American journal of human genetics* **99**, 40–55, <https://doi.org/10.1016/j.ajhg.2016.05.005> (2016).
69. Turcot-Dubois, A. L. *et al.* Long-term evolution of the CAZY glycosyltransferase 6 (ABO) gene family from fishes to mammals—a birth-and-death evolution model. *Glycobiology* **17**, 516–528, <https://doi.org/10.1093/glycob/cwm016> (2007).
70. Durán Aguilar, M. *et al.* Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. *Journal of Animal Breeding and Genetics* **134**, 49–59, <https://doi.org/10.1111/jbg.12238> (2017).
71. Lindholm-Perry, A. K., Kuehn, L. A., Freely, H. C. & Snelling, W. M. Genetic markers that influence feed efficiency phenotypes also affect cattle temperament as measured by flight speed. *Animal genetics* **46**, 60–64, <https://doi.org/10.1111/age.12244> (2015).
72. James, V. M. *et al.* Novel missense mutations in the glycine receptor beta subunit gene (GLRB) in startle disease. *Neurobiology of disease* **52**, 137–149, <https://doi.org/10.1016/j.nbd.2012.12.001> (2013).
73. Chabot, B., Stephenson, D. A., Chapman, V. M., Besmer, P. & Bernstein, A. The proto-oncogene c-kit encoding a transmembrane tyrosine kinase receptor maps to the mouse W locus. *Nature* **335**, 88–89, <https://doi.org/10.1038/335088a0> (1988).
74. Rieder, S. Molecular tests for coat colours in horses. *Journal of animal breeding and genetics = Zeitschrift für Tierzucht und Zuchtungsbiologie* **126**, 415–424, <https://doi.org/10.1111/j.1439-0388.2009.00832.x> (2009).
75. Schmutz, S. M. & Berryere, T. G. Genes affecting coat colour and pattern in domestic dogs: a review. *Animal genetics* **38**, 539–549, <https://doi.org/10.1111/j.1365-2052.2007.01664.x> (2007).
76. Szczerbal, I. *et al.* Ectopic position of duplicated KIT gene in African Nguni cattle, associated with color sidedness, confirms its shared ancestry with the Bos taurus lineage. *Animal genetics* **48**, 122–123, <https://doi.org/10.1111/age.12495> (2017).
77. Andersson, L. Genetic dissection of phenotypic diversity in farm animals. *Nature reviews. Genetics* **2**, 130–138, <https://doi.org/10.1038/35052563> (2001).
78. Rubin, C. J. *et al.* Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 19529–19536, <https://doi.org/10.1073/pnas.1217149109> (2012).
79. Carter, N. P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics* **39**, S16–S21 (2007).
80. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics*, elp017 (2009).
81. Xu, L., Hou, Y., Bickhart, D. M., Song, J. & Liu, G. E. Comparative analysis of CNV calling algorithms: literature survey and a case study using bovine high-density SNP data. *Microarrays* **2**, 171–185 (2013).
82. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559–575, <https://doi.org/10.1086/519795> (2007).
83. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* **17**, 1665–1674 (2007).
84. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
85. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639–1645, <https://doi.org/10.1101/gr.092759.109> (2009).
86. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols* **8**, 1551–1566 (2013).
87. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454, <https://doi.org/10.1038/nature05329> (2006).
88. Akey, J. M. *et al.* Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1160–1165, <https://doi.org/10.1073/pnas.0909918107> (2010).
89. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC bioinformatics* **12**, 35, <https://doi.org/10.1186/1471-2105-12-35> (2011).

## Acknowledgements

This study was supported by the National High Technology Research and Development Program of China (863 Program 2013AA102505-4), Cattle Breeding Innovative Research Team (cxgc-ias-03). This work was also supported in part by Agricultural Science and Technology Innovation Program of China (ASTIP-IAS-TS-9 and ASTIP-IAS-TS-16). L.Y.X. was supported by the Elite Youth Program in Chinese Academy of Agricultural Sciences.

## Author Contributions

L.Y.X. and L.Y. drafted the manuscript and performed the experiments. B.Z., H.N., J.M., X.P.S., M.Z., Y.C., X.G., L.L. and L.P.Z. collected the blood samples and extracted DNA samples. L.Y., L.Y.X., W.G.Z. and H.J.G. performed data analysis. L.Y.X. and J.Y.L. conceived of the study, J.Y.L. and G.E.L. participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-017-14768-0>.

**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