PLOS ONE

RESEARCH ARTICLE

Whole-genome sequencing reveals mutational landscape underlying phenotypic differences between two widespread Chinese cattle breeds

Yao Xu[1,2], Yu Jiang[1], Tao Shi[1], Hanfang Cai[1], Xianyong Lan[1], Xin Zhao[1], Martin Plath[1], Hong Chen[1]*

1 College of Animal Science and Technology, Northwest A & F University, Shaanxi Key Laboratory of Molecular Biology for Agriculture, Yangling, Shaanxi, China, **2** Institute of Biology and Medicine, College of Life Science and Health, Wuhan University of Science and Technology, Wuhan, Hubei, China

* chenhong1212@263.net

Citation: Xu Y, Jiang Y, Shi T, Cai H, Lan X, Zhao X, et al. (2017) Whole-genome sequencing reveals mutational landscape underlying phenotypic differences between two widespread Chinese cattle breeds. PLoS ONE 12(8): e0183921. https://doi.org/10.1371/journal.pone.0183921

Editor: Marinus F.W. te Pas, Wageningen UR Livestock Research, NETHERLANDS

Received: March 26, 2017

Accepted: August 10, 2017

Published: August 25, 2017

Data Availability Statement: All data in this study has been deposited in GenBank SRA database under accession number of SRP049655.

Funding: This study was supported by the National Natural Science Foundation of China (Grant No. 30972080, 31272408, 31600617), the Program of the National Beef Cattle Industrial Technology System (CARS-38), the National 863 Program of China (Grant No. 2013AA102505), and the Natural Science Foundation of Jiangsu Province (BK2011206). The funders had no role in study

Abstract

Whole-genome sequencing provides a powerful tool to obtain more genetic variability that could produce a range of benefits for cattle breeding industry. Nanyang (*Bos indicus*) and Qinchuan (*Bos taurus*) are two important Chinese indigenous cattle breeds with distinct phenotypes. To identify the genetic characteristics responsible for variation in phenotypes between the two breeds, in the present study, we for the first time sequenced the genomes of four Nanyang and four Qinchuan cattle with 10 to 12 fold on average of 97.86% and 98.98% coverage of genomes, respectively. Comparison with the Bos_taurus_UMD_3.1 reference assembly yielded 9,010,096 SNPs for Nanyang, and 6,965,062 for Qinchuan cattle, 51% and 29% of which were novel SNPs, respectively. A total of 154,934 and 115,032 small indels (1 to 3 bp) were found in the Nanyang and Qinchuan genomes, respectively. The SNP and indel distribution revealed that Nanyang showed a genetically high diversity as compared to Qinchuan cattle. Furthermore, a total of 2,907 putative cases of copy number variation (CNV) were identified by aligning Nanyang to Qinchuan genome, 783 of which (27%) encompassed the coding regions of 495 functional genes. The gene ontology (GO) analysis revealed that many CNV genes were enriched in the immune system and environment adaptability. Among several CNV genes related to lipid transport and fat metabolism, Lepin receptor gene (*LEPR*) overlapping with CNV_1815 showed remarkably higher copy number in Qinchuan than Nanyang ($\log_2$ (ratio) = -2.34988; *P* value = 1.53E-102). Further qPCR and association analysis investigated that the copy number of the *LEPR* gene presented positive correlations with transcriptional expression and phenotypic traits, suggesting the *LEPR* CNV may contribute to the higher fat deposition in muscles of Qinchuan cattle. Our findings provide evidence that the distinct phenotypes of Nanyang and Qinchuan breeds may be due to the different genetic variations including SNPs, indels and CNV.

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Integrating phenotypic variability with genomic variation is pivotal in both fundamental and applied biological sciences[1,2]. Most phenotypic traits, including naturally selected traits in wild populations[3] and those under artificial selection in domesticated animals[4], show quantitative genetic inheritance, and have complex genetic architectures[5,6]. Considerable progress has been made through high throughput sequencing to obtain whole genome sequences, which offer extensively promising efficient approaches for screening the molecular targets of selection and speciation[7,8].

Cattle are important farm animals and provide a major source of protein and fat for human populations worldwide, and thus variable traits related to the growth, carcass and meat quality are of vital interest in large-scale breeding projects of cattle[9]. Since the bovine genome was firstly sequenced from an inbred Hereford cow and her sire by capillary sequencing[10], the researches of resequencing and assembling genomes from different cattle breeds have quickly progressed. The whole-genome resequencing was firstly performed in a single Germany Fleckvieh bull, and 2.44 million single nucleotide polymorphisms (SNPs) and 115,000 small indels were identified by aligning to the reference assembly genome[11]. Kawahara-Miki et al.[12] resequenced the whole genome of a Japanese *Kuchinoshima-Ushi* cattle and revealed a total of 11,713 non-synonymous SNPs in protein-coding regions of 4,643 genes, and further phylogenetic analysis showed that the genetic background of *Kuchinoshima-Ushi* is quite distinct from the previously sequenced European domestic cattle breeds. The whole genome of a Korean Hanwoo bull was sequenced at a high coverage (45.6-fold), and 25 genes associated with meat quality and disease resistance were determined in the homozygous regions[13]. Moreover, gene copy number variations (CNVs) have been identified systematically at genome-wide level in cattle. Bickhart et al.[14] reported 1,265 CNV regions in six individuals from four American cattle breeds using high throughput sequencing. Additionally, the genome resequencing of two North American breeds, Black Angus and Holstein, revealed 790 putative CNV regions, which could be considered to be promising genetic markers for the identification across the genomes[15].

Even though genomic variability of traditional breeds is of vital interest from an agro-economic perspective as their traits could be desirable in primary breeding programs[2], we are still far from having a comprehensive understanding of genomic variations of different cattle breeds worldwide. For example, whole genome of the most widespread Chinese native cattle breeds has not been sequenced. Therefore, our study was designed to select a maximally informative subset of genomic markers for complex traits of agro-economic importance Similar to the case of natural selection, allele frequencies of loci underlying traits are expected to increase during the domestication process[16]. Therefore, comparing the whole genome sequences from distinct populations or breeds may contribute to identify the causative loci that affect phenotypic traits shaped by man-made selection.

We focused on two phenotypically distinct cattle breeds, Nanyang and Qinchuan, which belong to the native thoroughbred stock in China. Nanyang breed mainly originated from *Bos indicus* and was firstly domesticated in 1950. Qinchuan breed, known as *Bos taurus* in 6000~7000 B.C., and until 126 B.C. a little bloodline of *Bos indicus* was introduced into the breed and formed the domestic Qinchuan cattle. In general, Nanyang breed is characterized as a high acromion and narrow hindquarters, and have an average 380 kg of body weight for adult individuals, while Qinchuan has a much larger body weight (approximately 600 kg for adult male and 400 kg for female) and a typical thriving dewlap. In addition, as a well-known beef cattle breed in China, Qinchuan have higher meat production and quality than Nanyang due to its high marbling levels. We hypothesized that the phenotypic variability may be caused

by different genetic background between the two breeds. However, the whole genomic information of Nanyang and Qinchuan are unavailable so far, and comparative analyses of SNPs, indels and CNVs have not been feasible. In this study, we firstly sequenced the whole genomes of Nanyang and Qinchuan cattle by high throughput sequencing, and the results not only provide novel insight into the genetic difference at a whole-genome scale but also identify genomic loci that may be of vital interest in the programs for cattle breeding.

Materials and methods

Ethics statement

This study was approved by the Review Committee for the Use of Animal Subjects of Northwest A&F University. All animal experiments were carried out in strict accordance with institutional and state guidelines for animal care and all efforts were made to minimize suffering.

Sample preparation

Two famous Chinese cattle breeds (Nanyang and Qinchuan) were selected in our sequencing strategy. Blood samples (5 mL of each containing 20 U/mL heparin) were obtained from four Nanyang bulls and four Qinchuan bulls, which aged 4 years old and reared in the elite reservation farms from the Henan and Shaanxi province, respectively. All cattle were raised on a corn-corn silage diet. Genomic DNA (gDNA) was extracted from the whole blood with a QIAamp DNA Blood Maxi Kit (QIAgen, Valencia, CA, USA) according to manufacturer's instructions.

The adult (4 years old) Qinchuan cattle were slaughtered for tissue sampling in a commercial slaughterhouse located in the city of Yangling, under the supervision of the Animal Ethics Committee (AEC) and Northwest A&F University. Samples (fat, skeletal muscle, spleen, kidney, intestine, liver, heart, and lung) were immediately submerged in liquid nitrogen within 25 minutes of slaughter. In addition, fat from 16 cattle and skeletal muscle from 20 cattle were collected for total RNA and gDNA isolation. Total RNA was isolated with Trizol reagent (Takara, Liaoning, China) according to the manufacturer's protocol. The gDNA was also extracted from the fat and skeletal muscle (10 mg) by two rounds of proteinase K digestion and phenol-chloroform extraction. Further association analysis of CNV locus with phenotypic traits were conducted in a total of 191 Qinchuan cattle, and the traits were including body weight, body height, body length, heart girth, chest width, chest depth, height at hip cross, hucklebone width, hip width and rump length. Blood samples were collected and gDNA was isolated according to the procedures.

Preparation of fragment libraries and high-throughput sequencing

Libraries were prepared according to Illumina protocols. Each gDNA was sheared, polished and purified using minor modifications of the original Illumina Sample Preparation kit (Illumina 2006). Briefly, 10 μg gDNA was fragmented by nebulization for 5 min at a pressure of 32 psi N_2, and the sheared fragments were purified and concentrated using a QIAquick PCR purification spin column. To repair damaged gDNA ends and obtain 5'-phosphorylated blunt-ends (5'P), the fragments were end-repaired with T4 DNA polymerase, T4 phosphonucleotide kinase (PNK) and the Klenow fragment of *Escherichia coli* DNA polymerase. Terminal (3') A-residues were added following a brief incubation with dATP and Klenow 3'-5'exo-. Fragments were then ligated with solexa adaptors provided by the manufacturer. Adaptor-ligated fragments in the range of ~150–300 bp were selected using agarose electrophoresis. These small insert libraries were amplified independently using 18 rounds of PCR amplification and

standard primers PE 1.0 and PE 2.0 supplied by Illumina. After spin column extraction and quantitation, libraries were mixed (multiplex) at equimolar ratios to yield a total oligonucleotide mix of 10 nM.

Aliquots of multiplex libraries (5 pmol) were loaded onto the cluster generation station for single molecule bridge amplification on slides containing attached primers. The slide with amplified clusters was then subjected to step-wise sequencing using four-color labeled nucleotides on the Illumina 1G sequencing system for 36 cycles, which produces a theoretical fixed read length of 36 bp.

Sequence alignment and mapping

Paired-end sequence reads from Nanyang and Qinchuan were mapped against bovine genome assembly Bos_taurus_UMD_3.1, which was downloaded from the NCBI database. In this study, sequence scaffolds not yet assigned to specific chromosomes were excluded and no repeat masker was applied to the assembly.

The BWA algorithm ver. 0.5.0 (http://bio-bwa.sourceforge.net/bwa.shtml) was used for sequence alignment and consensus assembly. To obtain reliable alignment hits, the main parameters were defined for mapping, for instance, two mismatches were allowed between the read and the reference (mismatch penalty, misMsc = 2) when the sequence length < 60 bp, while it would be three mismatches if the sequence length > 60 bp. The sequence reads were not aligned with the inserting gaps, thus the parameter for read trimming (trimQual) was designated as 0. Moreover, the remaining indices were set according to the BWA default values. After read mapping, we discarded the reads mapped to multiple chromosomal positions and unmapped reads. Only reads with unique ungapped alignment were used for consensus calling and SNP detection.

SNP identification and annotation

Alignments of the reads from the Nanyang and Qinchuan were processed using SAMtools (http://samtools.sourceforge.net/samtools.shtml) to filter and report high quality SNP positions. SNP detection was performed using SAMtools 'pileup' command at default settings appropriate for diploid organisms. SNP filtering was performed in the following restricted conditions. The low-quality data were discarded (five bases with Q score < 20), and a minimum of 2× coverage depth (3× for the heterozygosity) was allowed for the initial identification of putative SNPs using SAMtools 'varFilter' command specifying fairly permissive minimum quality cutoffs. In addition, the heterozygous and homozygous SNPs were distinguished using an 80% cutoff of percent aligned reads calling the SNP. Consensus sequences for the SNP positions were generated using another tool from which SAMtools (and BWA) inherited code: MAQ[17].

The SNPs annotations were based on the 313,678 Bos taurus RefSeq in NCBI database. The cattle RefSeqs were aligned against Bos_taurus_UMD_3.1 using BLAT with the 'fine' option to obtain the genomic positions of genes, introns, and coding regions. In total, 63,213 RefSeqs were aligned against the reference genome. Among the aligned RefSeqs, the sequences with $>90\%$ coverage and a $<1\%$ error rate were selected. The selected genes were used to predefine the annotation data of all possible variants and pre-calculate the SIFT [18] predictions and scores. We selected the non-synonymous and frame shift (NS/F) that showed SIFT scores of <0.05 as the potentially damaging mutations.

Small indels and CNVs identification

A list of putative indels was generated for the two breeds from the paired-end reads, by combining the analysis of the algorithm BreakDancer [19] (http://breakdancer.sourceforge.net/breakdancermax.html) and the Pindel [20] program (http://trac.nbic.nl/pindel/), which computes the precise break points as well as the fragments inserted or deleted compared to the reference genome. In the preprocessing step, the BWA was first applied for mapping all the reads to the reference genome, and then the aligned results were examined to keep those paired reads that only one end can be mapped. For each paired read, the mapped end must be uniquely located in the genome without mismatch bases while the other end cannot be mapped to anywhere in the genome under a given threshold alignment score (20 to 36 bp reads). In the present study, for identifying small indels, the lenghth of 1 to 3 bp indels were obtained by setting relevant parameters.

The CNVs calls on the cattle genomes were identified using the CNV-seq program. Briefly, the Nanyang and Qinchuan reads were mapped to the Bos_taurus_UMD_3.1 using the BWA, and the output of BAM was converted into the "best-hit" format required by CNV-seq using the best-hit.Solexa.pl Perl script. Four consecutive sliding windows exhibiting a significant difference of read depth (minimum-windows-required = 4) enabled us to classify the region as a CNV. The Perl script was run with the default threshold values (*P*-value = 0.001 and $\log_2$ threshold = 1) and a window size setting of 2, to generate the putative CNVs from the best-hit files.

Quantitative PCR validation

Quantitative real-time PCR (qPCR) was performed to validate the CNVs from the whole genomes of Nanyang and Qinchuan cattle. Primers were designed *in silico* for 10 genic and 10 non-genic CNVs (S7 Table). Amplification efficiency for all primers was measured with different dilutions of gDNA (0.005 ng, 0.05 ng, 0.5 ng, 5 ng, 50 ng, 500 ng). Correlation coefficients (R^2), derived from linear regressions, ranged between 0.951 and 0.990, whereas the slope was between -3.164 and -3.842, indicating nearly 100% PCR efficiency for all the loci. The qPCR experiments were conducted in triplicate reactions, and each with a reaction volume of 20 μl containing 100 ng of gDNA, 10 μl SYBR® Premix Ex Taq TM II (Takara, Liaoning, China), and 10 pmol of primers. All reactions were amplified on a Bio-Rad CFX 96™ Real Time Detection System (Bio-Rad, Hercules, CA, USA). Gene relative expression was normalized to the expression of bovine glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene. Accordingly, the *LEPR* gene expression levels were quantified using Gene Expression Macro software (Applied Biosystems) by employing an optimized comparative Ct (ΔΔCt) value method, commonly designated as $2^{-\Delta\Delta Ct}$. In addition, the bovine *BTF3* gene was chosen as the diploid internal reference gene for the genomic qPCR analysis, and the ΔCt for each sample was calculated normalizing to *BTF3*[21], In addition, the copy number was confirmed based on the assumption that there were two copies of the DNA segment in the calibrator animals.

CNV annotation and gene ontology analysis

Gene content of cattle CNVs was assessed using Ensembl genes (ftp://ftp.ensembl.org/pub/current_fasta/bos_taurus/pep/). A total of 47,100 bovine peptides were retrieved from the Ensembl (Bos_taurus.UMD3.1.75.pep.abinitio). The canonical transcript record for each CNV gene was used to obtain a specific Ensembl protein ID, and the GO terms associated with the overlapping genes were analyzed using the agriGO server's Singular Enrichment Analysis (SEA) tool [22]. The significance of term enrichments were under or overrepresented in CNVs after Bonferroni correction.

Association analysis

SPSS v20.0 software (SPSS, Chicago, IL, USA) was used to analyze the associations of *LEPR* CNV with phenotypic traits in Qinchuan cattle by the One-way ANOVA method, and the relative copy number of *LEPR* was fitted as a continuous variable. Effects associated with farm, sex and season of birth (spring versus fall) were not into linear model, as the preliminary statistical analyses indicated that these effects did not have significant influence on variability of traits in the tested breeds. Thus the following model is used: $Y_{ijk} = \mu + A_i + CNV_j + e_{ijk}$, where Y_{ijk} is the observation of the growth traits; μ is the overall mean of each trait, A_i is the effect due to ith age, CNV_j is the fixed effect of jth CNVs type of *LEPR* gene and e_{ijk} is the random residual error.

Results

Whole-genome sequencing and mapping

We used Illumina technology to sequence gDNA from four Nanyang and four Qinchuan bulls, which were covered with an average mapping depth of 10 to 12 fold, respectively (Table 1). In total, approximately 333,930,957 reads (149,589,163 for Nanyang and 184,341,794 for Qinchuan) comprising 67.45 Gbp were generated, and values of quality ≥ 20 (Q_{20}) reached 100% (S1 Table). The obtained reads were mapped to the reference sequence (Bos_taurus_UMD_3.1) using the BWA algorithm[17], 96% of reads were successfully mapped to unique positions on the reference genome. On average, 97.86% (Nanyang) and 98.98% (Qinchuan) of bovine chromosome sequences were covered in our present study (Table 1; S1 Fig). Raw sequencing data in this study has been deposited in GenBank SRA database under accession number of SRP049655.

SNP/Indel annotation and comparison of two cattle breeds

We used SAMtools[23] to identify putative SNPs while mapping the aligned reads to the reference assembly. SNPs with quality values < 20 and a sequencing depth < 3 were filtered and discarded, resulting in final sets of 9,010,096 and 6,965,062 SNPs for the genomes of Nanyang and Qinchuan, respectively. 28.59% of all SNPs were transversions and 71.41% were transitions; 34.02% were homozygous and 65.98% heterozygous, with a ratio of 1:1.94 (S2 Table). To assess the false negative rates of SNPs in both breeds, the SNP list was compared to the locus obtained using Sanger sequencing assay. For Nanyang, a set of 517 SNPs were selected and validated by sequencing, and of these, 464 (89.7%) were identified as SNPs. Based on these results we calculated the false positive rate for SNP detection in Nanyang as $(1–464/517) * 100 = 10.3\%$. Additionally, the false positive rate for SNP validation of Qinchuan was calculated as $(1–470/502) * 100 = 6.4\%$.

We compared the identified SNP sets with those already published in the cattle dbSNP database (Build 133; ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/cow_9913/chr_rpts/). The SNPs of Nanyang had 4,429,836 (49.17%) positions overlapping with the cattle dbSNP database, whereas the Qinchuan shared 4,942,730 (70.98%) SNPs with the dbSNP database.

Table 1. Average coverage of the Nanyang and Qinchuan genomes.

Genome	Total reads	Mapped (%)	Identity (%)	Depth	Coverage[a] (%)
Nanyang	368,683,586	96.45%	99.40%	11.76	98.98%
Qinchuan	299,178,320	95.55%	99.24%	9.37	97.86%

[a]Fold coverage was calculated by aligning the assembling sequences to the reference chromosomes (Bos_taurus_UMD_3.1) used for mapping.

https://doi.org/10.1371/journal.pone.0183921.t001

Distributions of the remaining novel SNPs on each of the 30 chromosomes were depicted in S2 Fig. Using Pindel and BreakDancer, 154,934 and 115,032 small indels of 1 to 3 bp length were identified in the Nanyang and Qinchuan, respectively. The frequencies of three indel types (insertions, deletions, and insertions within deletions) in each chromosome were shown in S3 Table. Indels of 1 bp length prevail in frequency (S3 Fig), which was in accordance with a previous study by Kawahara-Miki et al. [12]. In addition, we found that the SNPs and small indels in each chromosome of Nanyang were more than that in Qinchuan (S4 Fig). All SNPs and indels in this study have been submitted to the dbSNP at GenBank under the handle NWAF_LMBA.

All detected SNPs and indels in Nanyang and Qinchuan were annotated on the basis of the NCBI Reference Sequence Database (RefSeq: ref_Bos_taurus_UMD_ 3.1_ gnomon_top_level. gff3). Most of the SNPs and small indels were located in intergenic regions or introns, and only 0.84% were located in exons (Fig 1). Of these SNPs in coding regions of the Nanyang, 37,309 (49.68%) were non-synonymous substitutions, including missense and nonsense mutations, which were distributed in 11,712 functional genes. Additionally, a total of 355 small indels were located in coding regions, 239 of which (67.32%) were identified as variations that may cause gains or losses of stop codons by shifting the open reading frame (ORF); these occurred in 226 genes. By contrast, 30,389 non-synonymous SNPs (50.37%) were found in 10,991 genes in Qinchuan cattle, and 231 indels (71.52%) in 215 genes led to frame-shifts (S4 Table).

Non-synonymous SNPs, frame-shift, and indels (NS/FS/Indel) within a coding DNA sequence can affect the expression and function of important genes. Thus, a list of genes with more than 100 SNPs altogether, or more than 50 NS/FS/Indel, was compiled in Nanyang and Qinchuan genomes. Comparison between the two breeds revealed that the Nanyang genome had 32 NS/FS/Indel genes, while the Qinchuan had only 25 NS/FS/Indel genes. Among those

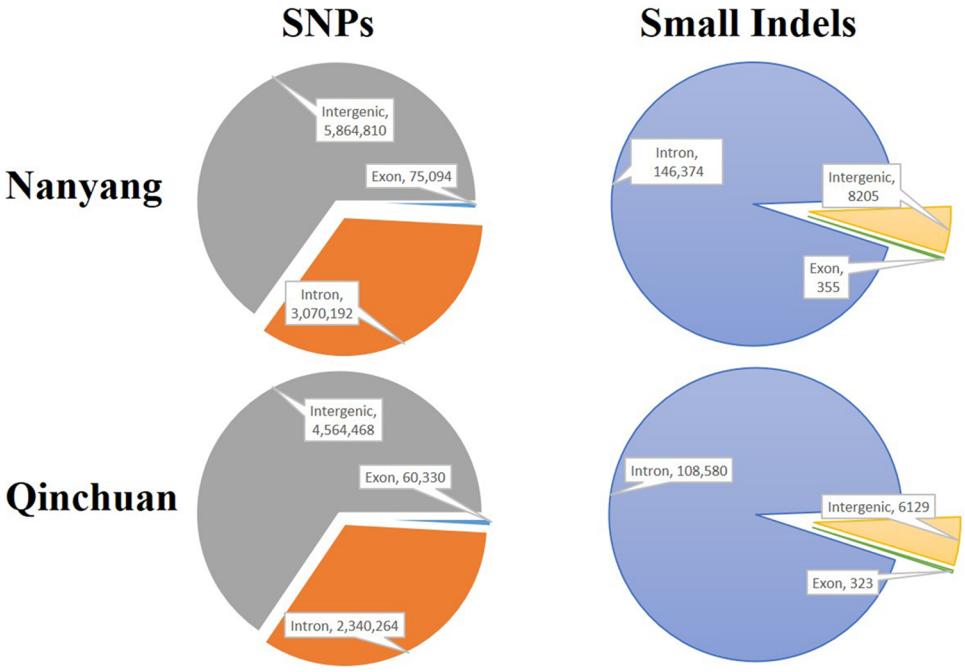

Fig 1. The number of SNPs and small indels distributed in the Nanyang and Qinchuan genomes (exon, intron, and intergenic).

https://doi.org/10.1371/journal.pone.0183921.g001

genes, the obscurin-like 1 (*OBSL1*) gene, a novel cytoskeletal protein related to obscurin [24], showed more NS/FS/Indel in Nanyang than Qinchuan (102 *vs*. 78). The *OBSL1* gene has a larger genome span (22.091 kb) and transcript (6.046 kb) compared to several other functional genes. Importantly, two splice variants of the *OBSL1* gene (ENSBTAP00000053524 and ENSBTAT00000060992) were located on chromosome 2, and 42 mutations (NS/FS/Indel) in the two variants were recorded in Ensembl.

Copy number variation and validation through qPCR

Read sets of Nanyang and Qinchuan were conducted by aligning to the Hereford reference sequence assembly (Bos_taurus_UMD_3.1), while copy number variations (CNVs) were identified by compring the significantly different regions between the Nanyang and Qinchuan (control sample) mapped read sets. Overall, we detected 2,907 cases of CNV, amounting to approximately 9.9 Mbp and corresponding to 0.37% of the cattle genome (Table 2). The

Table 2. The detailed information of CNVs in our study.

Chr	Chromosome length	No. CNV	Total CNV length[a]	% length in CNV	% No. CNV	Max length	Mean length	Median length
1	158337067	99	277056	0.1750	3.406	10212	2799	2220
2	137060424	160	606060	0.4422	5.504	22644	3788	2664
3	121430405	110	376956	0.3104	3.784	48396	3427	2220
4	120829699	105	306360	0.2535	3.612	22644	2918	2220
5	121191424	172	808968	0.6675	5.917	158952	4703	2220
6	119458736	84	202908	0.1699	2.890	11988	2416	1776
7	112638659	113	393828	0.3496	3.887	24420	3485	2220
8	113384836	98	284160	0.2506	3.371	22200	2900	2220
9	105708250	113	317904	0.3007	3.887	12432	2813	2220
10	104305016	119	391164	0.3750	4.094	25308	3287	2220
11	107310763	47	121212	0.1130	1.617	9324	2579	2220
12	91163125	377	1499388	1.6447	12.969	52836	3977	2664
13	84240350	48	121212	0.1439	1.651	6216	2525	1998
14	84648390	49	127872	0.1511	1.686	10656	2610	1776
15	85296676	116	366300	0.4294	3.990	22200	3158	2220
16	81724687	37	103008	0.1260	1.273	22644	2784	1776
17	75158596	90	291708	0.3881	3.096	13764	3241	2220
18	66004023	89	267732	0.4056	3.062	10212	3008	2220
19	64057457	47	188700	0.2946	1.617	25752	4015	2220
20	72042655	28	68820	0.0955	0.963	4884	2458	1998
21	71599096	66	196692	0.2747	2.270	14652	2980	2220
22	61435874	14	33300	0.0542	0.482	4440	2379	2220
23	52530062	54	182928	0.3482	1.858	9768	3388	2664
24	62714930	29	147408	0.2350	0.998	64380	5083	2220
25	42904170	12	31524	0.0735	0.413	7548	2627	1998
26	51681464	29	88800	0.1718	0.998	17760	3062	2220
27	45407902	48	154956	0.3413	1.651	11988	3228	2220
28	46312546	24	55056	0.1189	0.826	5328	2294	1776
29	51505224	81	327228	0.6353	2.786	35076	4040	2220
X	148823899	449	1582860	1.0636	15.445	47952	3525	2220
Total	2660906405	2907	9922068	0.0037	-	158952	3183	2220

[a]The distribution and size characteristics of CNVs detected through comparison of the read sets mapped to the Bos_taurus_UMD_3.1 reference assembly.

https://doi.org/10.1371/journal.pone.0183921.t002

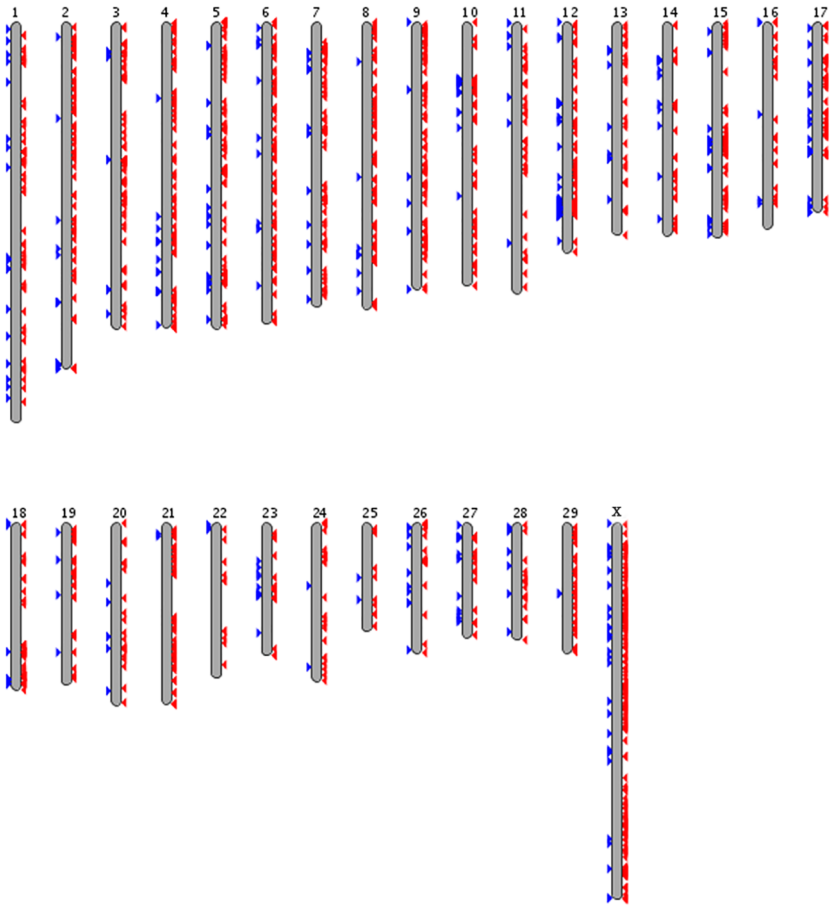

Fig 2. Schematic diagram of copy number variations (CNVs) regions in cattle genome. Blue arrowheads located on the chromosome ideograms represent copy number gains in Nanyang against with Qinchuan; while the red arrowheads represent copy number gains in Qinchuan against with Nanyang (Nanyang CNV losses). Note that several CNVs may combine and display as a single arrowhead due to their proximity in the genome.

https://doi.org/10.1371/journal.pone.0183921.g002

detected CNVs were unevenly distributed across chromosomes (Fig 2). Specifically, the percentage of CNVs numbers ranged from 0.413% to 5.917% on most other chromosomes, while 12.969% of CNVs (377) were located on chromosome 12, identifying chromosome 12 as the most polymorphic chromosome. The greatest number (449; 15.445%) of CNVs, however, was discovered on chromosome X, which is consistent with an earlier study [14]. Regions with CNVs varied in size between 1,776 bp and 158,952 bp, and the mean and median CNV length were 3,183 bp and 2,220 bp, respectively (Table 2). Detailed information on CNVs is provided in S5 Table.

To confirm the accuracy of CNV assessment from genome sequencing, quantitative PCR (qPCR) was conducted for a randomly selected subset of 20 CNVs, which were divided into two groups (10 genic and 10 non-genic CNVs). Both groups contained five gains (log_2_Ratio > 0.5) and five losses (log_2_Ratio < -0.5) of CNV status (Table 3). 90% of our qPCR results (i.e., 18 of 20 validated) agreed with the CNV predictions in these regions, and only CNVR_161 and CNVR_1410, harboring the *SSBP2* and *SNTA1* genes, respectively, were not congruent with the predictions from out genome sequencing results (Fig 3).

Table 3. The log_2 ratio and *P*-value of CNVs selected for qPCR validation.

CNV position	Ensembl gene	CNV type	log_2 ratio[a]	*P*-value[a]
Chr2_CNV_2113	ACTL8	gain	1.290	4.01E-87
Chr2_CNV_2062	RAPH1	gain	1.750	1.17E-45
Chr5_CNV_1452	LOC617219	gain	2.106	1.94E-112
Chr21_CNV_2724	GABRG3	gain	1.087	1.66E-20
ChrX_CNV_558	IL1RAPL2	gain	3.060	8.08E-59
Chr12_CNV_1028	-	gain	1.992	1.87E-39
Chr14_CNV_27	-	gain	2.484	2.52E-86
Chr15_CNV_392	-	gain	2.018	4.19E-43
Chr20_CNV_2034	-	gain	3.822	8.51E-184
Chr27_CNV_1684	-	gain	2.283	4.47E-168
Chr2_CNV_2150	COL5A2	loss	-1.278	1.88E-20
Chr3_CNV_1815	LEPR	loss	-2.350	1.53E-102
Chr4_CNV_1934	SHH	loss	-1.720	9.08E-45
Chr7_CNV_161	SSBP2	loss	-1.228	2.15E-32
Chr13_CNV_1410	SNTA1	loss	-1.393	4.28E-44
Chr7_CNV_136	-	loss	-1.108	4.69E-41
Chr11_CNV_505	-	loss	-1.907	2.89E-39
Chr13_CNV_1398	-	loss	-6.409	1.21E-101
Chr20_CNV_2047	-	loss	-2.976	1.18E-70
Chr22_CNV_7	-	loss	-3.015	4.01E-70

[a]The log_2 ratio and *P*-value were obtained from the CNV-seq software. Positive log_2 ratios represented that the copy number in Nanyang was higher than the Qinchuan, while the negative values indicated higher copy number in Qinchuan than Nanyang.

https://doi.org/10.1371/journal.pone.0183921.t003

To further evaluate CNV between breeds, the same 20 CNV regions were quantified in 10 Nanyang and 10 Qinchuan individuals. Relative copy numbers (CNs) were calculated based on the assumption that there were two copies of each DNA segment in the calibrator (the Qinchuan sequenced in our initial CNV survey). As illustrated in S5 Fig, average copy numbers of Nanyang were larger than those of Qinchuan breed in the group of CNV gains. In contrast, mean copy numbers of Nanyang were always lower than those of Qinchuan in the case of CNV losses.

Gene content and gene ontologies of cattle CNV regions

To identify potential functional roles associated with the putative CNVs, genes completely or partially overlapping with these CNVs were retrieved from Ensembl[25]. 27% (783/2907) of all CNVs involved 495 partially or entirely functional genes. Gene ontology (GO), a standardized gene classification system[26], was performed to determine the likely biological effects of CNV genes using the web-based tool agriGO [27]. Statistically significant over-representations were observed for multiple categories (Table 4), which showed that the GO terms "G-protein coupled receptor signaling pathway" (GO:0007186; $P<0.01$), "immune system process" (GO:0002376; $P<0.01$), "Olfactory receptor activity" (GO:0004984; $P<0.01$), "Integral to membrane" (GO:0016021; $P<0.01$), were enriched among the CNV genes in the Nanyang. The analyzed CNV genes, including ATP-binding cassette A13 (*ABCA13*), leptin receptor (*LEPR*) and attractin (*ATRN*), were consistent with earlier reports by Liu et al.[21] and Wang et al.[28]. The results imply that artificial selection may have driven the gain or loss of copy

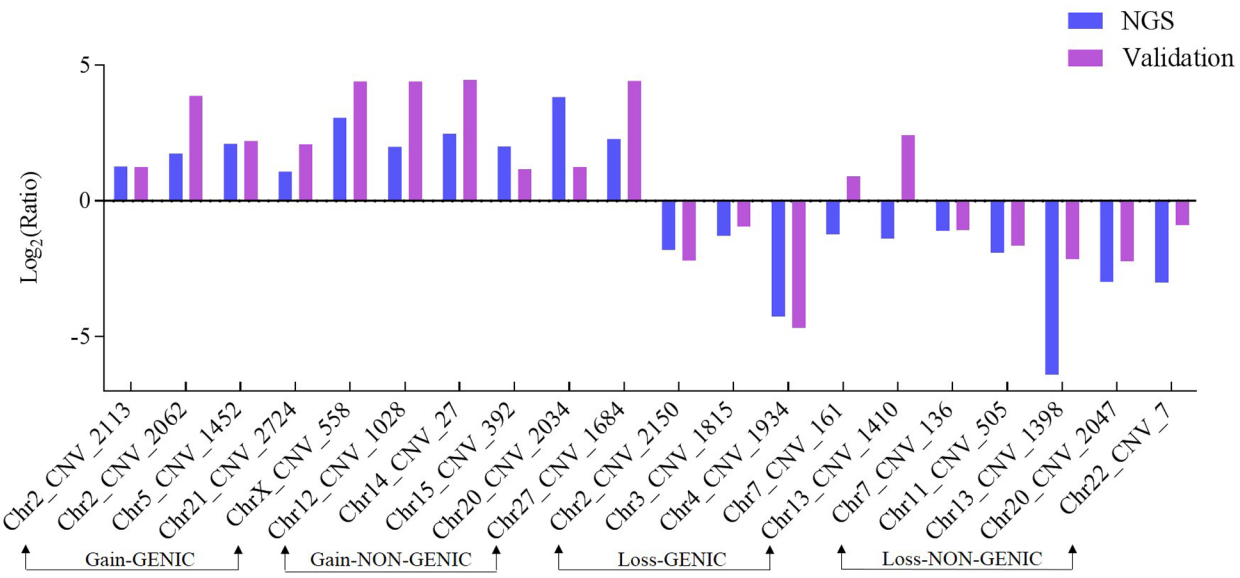

Fig 3. Validation of copy number variations (CNVs) detected from whole-genome resequencing using quantitative PCR. Validation results are shown for four CNVs groups: gain in genic, gain in non-genic, loss in genic, and loss in non-genic. Values on y-axis represented the $\log_2$ (2^Sample signal) for resequencing and $\log_2 2^{-\Delta\Delta Ct}$ for qPCR validation, respectively. The two values located in the same side of the "0" line represented the same CNV types, while the values located in two side represented the different CNV types.

https://doi.org/10.1371/journal.pone.0183921.g003

numbers and thus, specific gene dosages are necessary to form several breed-specific phenotypic characteristics.

Functional validation of copy number variation of *LEPR* gene

According to the above GO enrichment analysis, many CNV genes were over-represented in the terms related to environmental response, which is consistent with other studies of mammalian genomes [29,30]. However, in our study, it was intriguing to note that the leptin receptor (*LEPR*) gene was overlapped with a CNV region (CNV_1815) in chromosome 3 (Fig 4a and 4b). Sequence analysis revealed that this region (*LEPR* CNV) contained partial repetitive elements (LINEs and SINEs) and a small CpG island (Fig 4c), which play important roles in the RNA-mediated gene duplication and methylated regulation [31,32]. In addition, the CNV region showed significantly higher copy numbers in Qinchuan cattle compared with Nanyang, thus we hypothesized that the *LEPR* gene may contribute to higher meat quality of Qinchuan than Nanyang cattle due to its role in lipid metabolism[33].

To further test the hypothesis, we sought to determine whether copy number variation affects the expression level of *LEPR* gene in Qinchuan cattle. Firstly, expression profiling of the *LEPR* gene in different tissues revealed that *LEPR* was expressed at a high level in fat and at a moderate level in skeletal muscle tissue (S6 Fig). Thus, correlations between copy number and mRNA levels of the *LEPR* gene were examined in fat ($n = 16$) and skeletal muscle ($n = 20$) separately; results are shown in Fig 5. In both data sets, positive correlations were uncovered (linear regressions: $R^2 = 0.61$, $P = 0.0004$ for fat; $R^2 = 0.40$, $P = 0.0029$ for skeletal muscle). In addition, to evaluate whether the potential effects of CNV locus is causative for the different phenotypes between Nanyang and Qinchuan cattle, the association analysis of *LEPR* CNV with growth traits were conducted in Qinchuan population. In the CNV testing analysis, aberrant segments were identified as a CNV locus according to the $\log_2$ (ratio of test/control). Similarly, the copy number types of the *LEPR* were classified as gain (>0.5), loss (<-0.5) and median (<|±0.5|)

Table 4. Enriched gene ontology (GO) terms of genes in identified CNV regions (kolmogorov-smirnov *P*-value ≤ 0.05).

Ontology[a]	GO.ID	Term	KS *P*-value	Validated in previous study
CNV.GO_BP	GO:0050911	Detection of chemical stimulus involved in sensory perception of smell	< 1e-30	-
CNV. GO_BP	GO:0007186	G-protein coupled receptor signaling pathway	< 1e-30	Cattle (Liu. G. E. etal. 2010) Pig (Wang. J. etal 2012)
CNV.GO_BP	GO:0006334	Nucleosome assembly	8.50E-08	Pig (Wang. J. etal 2012)
CNV.GO_BP	GO:0050877	Neurological system process	0.00066	Cattle (Liu. G. E. etal. 2010) Pig (Wang. J. etal 2012)
CNV.GO_BP	GO:0002376	Immune system process	0.00081	Cattle (Bickhart. D. M. etal 2012)
CNV.GO_BP	GO:0007165	Signal transduction	0.0054	Cattle (Liu. G. E. etal. 2010)
CNV.GO_BP	GO:0019882	Antigen processing and presentation	0.01789	Pig (Wang. J. etal 2012)
CNV.GO_BP	GO:0007600	Sensory perception	0.03015	Cattle (Liu. G. E. etal. 2010)
CNV.GO_BP	GO:0051056	Regulation of small GTPase mediated signal transduction	0.03445	Pig (Wang. J. etal 2012)
CNV.GO_BP	GO:0007166	Cell surface receptor signaling pathway	0.03971	Cattle (Liu. G. E. etal. 2010) Pig (Wang. J. etal 2012)
CNV.GO_MF	GO:0004984	Olfactory receptor activity	< 1e-30	Cattle (Liu. G. E. etal. 2010) Pig (Wang. J. etal 2012)
CNV.GO_MF	GO:0004930	G-protein coupled receptor activity	< 1e-30	Cattle (Liu. G. E. etal. 2010)
CNV.GO_MF	GO:0046982	Protein heterodimerization activity	0.00736	Pig (Wang. J. etal 2012)
CNV.GO_MF	GO:0070330	Aromatase activity	0.00905	Pig (Wang. J. etal 2012)
CNV.GO_MF	GO:0003824	Catalytic activity	0.04975	Cattle (Bickhart. D. M. etal 2012)
CNV.GO_CC	GO:0016021	Integral to membrane	< 1e-30	Cattle (Liu. G. E. etal. 2010) Pig (Wang. J. etal 2012)
CNV.GO_CC	GO:0005886	Plasma membrane	< 1e-30	Cattle (Liu. G. E. etal. 2010)
CNV.GO_CC	GO:0000786	Nucleosome	5.70E-07	Pig (Wang. J. etal 2012)
CNV.GO_CC	GO:0044464	Cell part	0.03365	Cattle (Bickhart. D. M. etal 2012)
CNV.GO_CC	GO:0005623	Cell	0.03365	Cattle (Bickhart. D. M. etal 2012)

[a]The three Ontologies of GO enrichments BP, MF, and CC represented Biological Process, Molecular Function, and Cellular Component, respectively.

https://doi.org/10.1371/journal.pone.0183921.t004

based on the $\log_2 2^{-\Delta\Delta Ct}$ relative to the control sample by qPCR analysis. In the statistical model, the *LEPR* copy numbers were normalized to the sequenced Qinchuan cattle (control sample). Correspondingly, the copy number of the gain, loss and median was designated as ≥3, <2 and 2 copies, respectively. As shown in Table 5, the *LEPR* CNV was significantly associated with body weight, hucklebone width and rump length, the individuals with copy number gain had higher values than those with loss or median ($P < 0.05$), which was consistent with the correlation analysis of *LEPR* CNV and transcriptional level.

Discussion

The primary goal of our present study was to provide a comprehensive list of sequence variation at the whole genome scale for two widespread Chinese cattle breeds (Nanyang and Qinchuan) and to identify potential loci that might be related to phenotypic differences between the two breeds. In previous studies, examining whole sequence variation is often relied on SNP array[34], array comparative genomic hybridization (aCGH) [21,35] or exome sequencing [36]. However, the SNP array and aCGH approaches obviously have a limited sensitivity and are predisposed to high false positive/negative calling [37]. Exome sequencing provides limited information as the sequence information of regulatory and intergenic regions is unavailable [38]. Thus, in this study, next-generation sequencing, allowing sequence construction at a higher effective resolution and sensitivity[39], was used to identify polymorphisms from genomes of the investigated breeds.

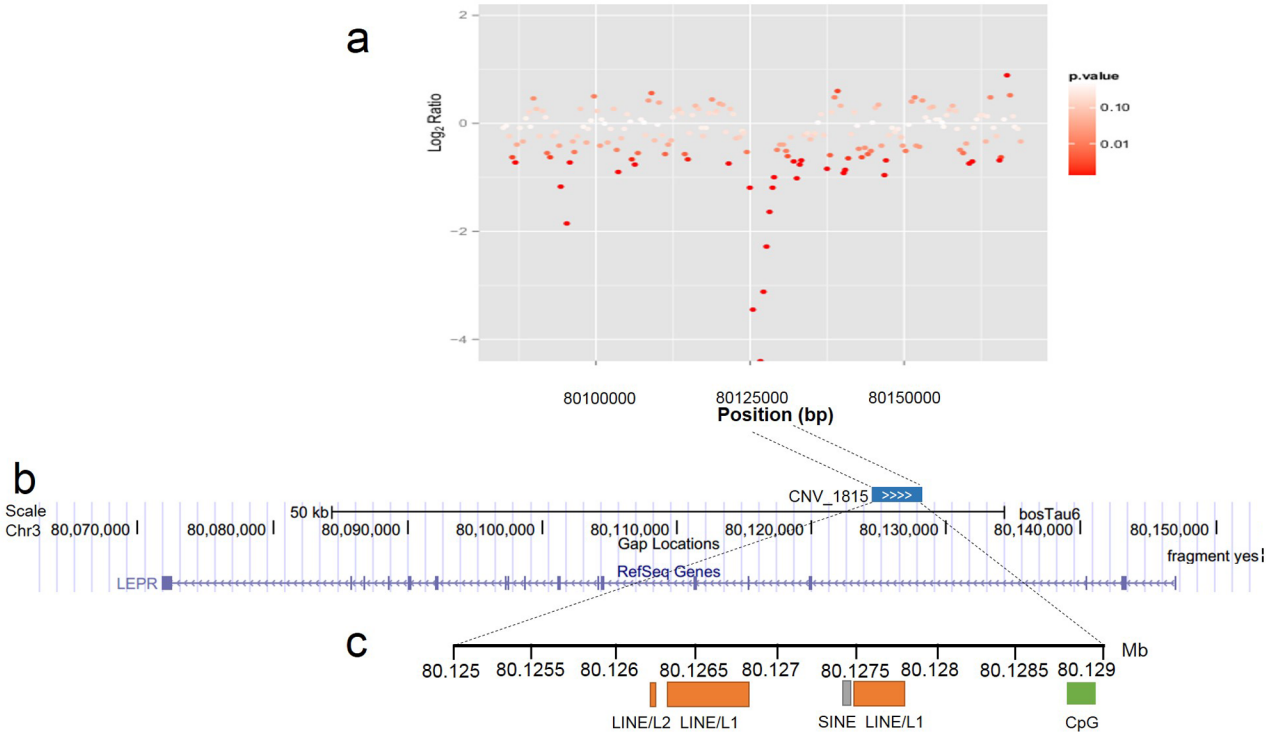

Fig 4. The bovine Lepin receptor (*LEPR*) gene overlapping with copy number variation (CNV) region. a represented $\log_2$ ratio plot of the *LEPR* gene region. Each point shows the $\log_2$ ratio of Nanyang reads mapped to Qinchuan reads. Points are coloured based on the $\log_{10}$ p-value calculated by the CNV-seq software. b represented the *LEPR* gene region as visualized using the UCSC Genome Browser. The precise boundary of the CNV_1815 that resides in this region is shown and labelled. c represented the localization and composition of important elements in the CNV_1815. The small bars in orange are partial LINEs, and the gray shows SINEs. A small CpG island is marked with a green bar.

https://doi.org/10.1371/journal.pone.0183921.g004

Phenotypic diversity of domesticated animals has been shaped by man-made selection, including selective sweeps, which ought to leave a signature in the genome of domesticated strains or breeds [40]. Our study was motivated by the idea that comparing breeds with contrasting phenotypes may provide molecular basis underlying phenotypic differences, and thus provide novel insights into the effects of artificial selection on the genomes of domesticated animals. The Nanyang and Qinchuan breeds presented marked differences in several phenotypic traits of agro-economic interest, for example, Qinchuan cattle have a higher meat quality and especially higher marbling grades than Nanyang cattle. In fact, our genomic sequencing approach identified a number of loci that could be related to these phenotypic differences. The sequencing results indicated that Nanyang showed more genetic variation than Qinchuan, including SNPs, small indels and CNVs, suggesting that Nanyang cattle could have a larger effective population size (*Nm*), while in case of Qinchuan, they may own fewer breeding bulls, leading to the loss of genetic variability due to drift and bottlenecks. In addition, Nanyang may possess higher divergence comparing to the *Bos taurus* reference sequence, which indicate that Nanyang has inherited the genetic characteristics from *Bos indicus*, and it may be phylogenetically more distinct from the reference cattle genome.

We sequenced the Nanyang and Qinchuan genomes at 10–12 fold depth, and the averages of 97.9% and 98.9% of the whole genome sequences were covered, respectively. We used a slightly higher read depth than a previous study on the Fleckvieh breed (7.4-fold), which uncovered 2.44 million SNPs [11]. Lee et al. [13] sequenced the genome with 45.6-fold depth

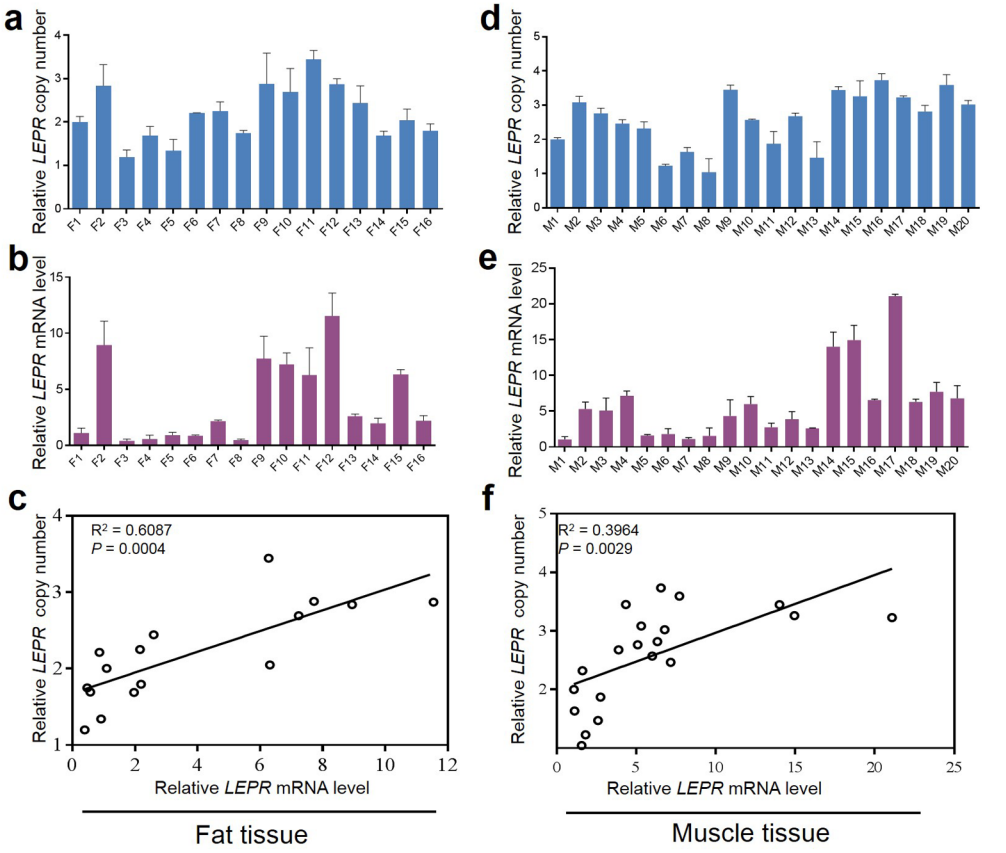

Fig 5. Relationship between the copy number variation (CNV) region of the Lepin receptor (*LEPR*) gene and transcript levels in fat and muscle tissue. a~c represented the correlation result in fat tissue (n = 16, F1~F16). The relative copy numbers and expression of *LEPR* CNV region were normalized as compared to that from the F1 individual. d~f represented the correlation result in muscle tissue (n = 18, M1~M20). The relative copy numbers and expression of *LEPR* CNV region were normalized as compared to that from the M1 individual.

https://doi.org/10.1371/journal.pone.0183921.g005

Table 5. Statistical association analysis of bovine *LEPR* gene copy number variations with phenotypic traits in Qinchuan cattle.

Growth traits	CNVs types[a] (Mean ± SE)			*P*-value
	Gain (n = 32)	Loss (n = 36)	Median (n = 123)	
Body weight, kg	408.06±18.93[a]	381.20±17.84[ab]	366.37±9.77[b]	0.043
Body height, cm	127.38±1.45	127.64±1.37	127.27±0.75	0.972
Body length, cm	138.19±2.98	133.89±2.81	131.58±1.54	0.144
Heart girth, cm	178.16±2.99	174.17±2.82	172.30±1.55	0.222
Chest width, cm	38.06±1.12	37.47±1.06	37.05±0.58	0.714
Chest depth, cm	62.50±1.12	61.33±1.05	61.73±0.58	0.740
Height at hip cross, cm	124.56±1.55	125.03±1.46	124.50±0.80	0.950
Hucklebone width, cm	24.41±1.04[a]	22.72±0.98[ab]	21.93±0.54[b]	0.037
Hip width, cm	42.56±0.87	40.72±0.82	41.57±0.45	0.311
Rump length, cm	44.75±0.73[a]	42.72±0.69[b]	42.98±0.38[b]	0.035

[a] The copy number types of the gain, loss and median was designated as ≥3, <2 and 2 copies, respectively. Values with different superscripts (a, b) within the same row differ significantly at $P < 0.05$.

https://doi.org/10.1371/journal.pone.0183921.t005

in the case of the Korean Hanwoo cattle; however only 4,781,758 SNPs were identified—by far less than in our present study (9,008,518 for Nanyang; 6,963,517 for Qinchuan). Sequencing Japanese Kuchinoshima-Ushi cattle uncovered 6,303,790 SNPs [12]. while 3,755,663 (19-fold) and 3,246,211 SNPs (22-fold) were found in Holstein and Black Angus (North American cattle), respectively [15]. These investigations suggested that the number of genomic SNPs showed remarkable differences among different cattle breeds, and the genomic variabilities may become valuable resources for future breeding campaigns—e.g, the international cattle industry my focus on desirable phenotypic traits as seen in Chinese breeds, like the famous meat structure and quality of Qinchuan cattle.

It is straight forward to predict that non-synonymous SNPs have a high predictive power to explain phenotypic differences[41], because non-synonymous substitutions lead to an altered protein structure and/or spatiotemporal patterns of gene expression [42]. In our study, the percentage of non-synonymous SNPs in coding regions was 41.14% and 40.13% for Nanyang and Qinchuan, respectively, which was higher than any previous report from whole-genome resequencing approaches in cattle [12,15]. Several genes with non-synonymous SNPs (equal or more than 10 non-synonymous SNPs per gene) have been reported to be associated with agro-economically important traits, such as growth rate and meat production [43]. Growth-related genes like growth hormone (*GH*) [44], growth hormone receptor (*GHR*) [43] and insulin-like growth factor 1 (*IGF1*) [45], comprised more non-synonymous SNPs in Nanyang than that in Qinchuan. Contrarily, Qinchuan genome showed a larger number of non-synonymous SNPs in genes associated with meat (i.e., muscle tissue) development, such as diacylglycerol O-acyltransferase 1(*DGAT1*) [46] and leptin (*LEP*) [47]. Future studies will need to be conducted to understand the functional roles of these non-synonymous substitutions during muscle formation and development in Qinchuan cattle—research that will be of interest for the international cattle industry seeking to improve meat quality of different breeds worldwide.

Moreover, a list of genes with > = 50 NS/FS/Indel were identified in this study, and most of the genes were important in cellular defense, adaptive immunity, and environmental response, which corresponded well with the previous studies in human and other animals [29,30,48,49]. For example, sequence variations of the major histocompatibility complex (MHC), a master coordinator of specificity in both adaptive and innate immune systems, are related to a large number of infectious, autoimmune and other diseases [50]. In addition, Bergen et al. [51] established the genome-wide association study and showed that SNPs of MHC were significantly associated with schizophrenia in a Swedish population.

A substantial number of CNVs were detected in the Nanyang when aligning it to the Qinchuan. Statistics analysis showed that the minimal and mean length of CNVs were 1,776 bp and 3,183 bp, respectively, which is in accordance with other sequencing-based studies on cattle [52], but considerably shorter than what has been reported based on SNP/CGH array methodology (several mega-base pairs) [21,53]. This discrepancy can be attributed, in part, to different criteria for reporting CNVs. The approach we used here can artificially break a single CNV into multiple CNVs [54]. We also compared our results to previous reports on CNVs in cattle, which were identified using two different technologies (S6 Table). Stothard et al. [15] applied whole-genome resequencing to detect CNVs in Holstein and found 790 cases of presumed CNV by mapping the Holstein genome to the Btau4.0 reference sequence. Only 11 out of those 790 CNVs (1.4%) were identical or overlapping with CNVs detected in our present study. Using next-generation sequencing approach, Bickhart et al. [14] reported 1,265 CNVs across all chromosomes in five cattle, and 5.1% (65/1265) overlapped with our results. This comparison suggests that we detected a great number of previously unknown cases of CNV in the cattle genome. Moreover, in order to validate the sequencing-based CNV call set, we selected 20 CNVs for qPCR and achieved a confirmed rate of 90%, which was higher than

most confirmation rates in previous studies, such as the study by Hou et al. [53] in 15 cattle (60.00%), Bickhart et al. [14] detected 12 CNV loci in BINE, BTAN1, BTAN2, BTAN3 and BTHO cattle (82.14%) but a little lower than that in modern domesticated cattle (91.67%) reported by Liu et al. [21]. Notably, the previous study has not detect breed-specific CNVs, and we for the first time demonstrated major differences in copy numbers of certain loci in randomly selected individuals from Nanyang and Qinchuan (n = 10 individuals each). The results were completely concordant with our sequencing data.

We detected a considerable number of annotated genes (495 Ensembl genes) in CNVs regions. As previously shown, genes in a CNV region can contribute to phenotypic variation by changing gene structure, alternating gene regulation, exposing recessive alleles, and other mechanisms [55]. In recent years, many functional CNV genes have been investigated in cattle, for instance, the *PLA2G2D* gene, related to milk production and meat quality, was highly duplicated in beef cattle breeds [15]. Xu et al. [56,57] reported that the duplicated bovine *MYH3* and *MICAL-L1* genes, located on the quantitative trait loci (QTLs) for body weight, were associated with growth traits in Chinese cattle. In our study, GO analysis of CNV genes revealed that terms such as "innate and adaptive immunity", and "receptor recognition" were enriched, which is consistent with investigations on CNVs in human, mouse, dog, and cattle [29,30,48,58]. Among several genes related to lipid transport and fat metabolism, the *LEPR* gene was explored in more detail as it showed a great deal of copy number differences between the Nanyang and Qinchuan breeds. The CNV region contains variable copies of the LINEs, SINEs and CpG elements, which could be relevant for the mechanism of action of this noncoding variation. Wright et al. [59] has reported that in chicken the copy number variations in intron 1 of *SOX5* were associated with the pea-comb phenotypes. The LEPR protein, a receptor of leptin, is produced by fat cells, and influences food intake, fat metabolism, and reproductive functions [60]. A previous study reported that an increased copy number in the "E2 DNA" region (exon-intron junction) of the *LEPR* gene lead to an increased fat deposition in humans, and *LEPR* CNV is thought to be involved in obesity and type 2 diabetes mellitus [61]; this suggests that artificial selection in cattle seems to have selected for a trait that in other biological systems occurs as a rare disease. By referring to the cattle QTLdb (quantitative trait loci database) [62], the bovine *LEPR* gene was mapped at an 84 cM interval on chromosome 3, in which the QTL region (no. 13,158) for the fat thickness at the 12th rib trait was located [43]. Additionally, significantly positive correlations were observed between the copy number of the *LEPR* CNV and its transcription levels in skeletal muscle and fat tissues, suggesting that the CNV region of the *LEPR* gene may indeed affect phenotypic traits of cattle by affecting the copy number of transcripts and ultimately, the LEPR protein. A straight forward working hypothesis arising from our study is that the higher copy number of the *LEPR* CNV in Qinchuan cattle is one of the main factors responsible for increased fat deposition in muscle tissue—a desirable trait for meat production.

Whole genome sequencing presented the first description of genomic variations in the Chinese Nanyang and Qinchuan cattle. In comparison with the studies previously reported, the two Chinese cattle genomes showed a higher degree of genetic diversity than those of other cattle breeds, and the Nanyang presented more abundant variations than Qinchuan. According to the GO enrichment analysis results, we conclude that the bovine *LEPR* gene may be one of the causative genes contributing to the different phenotypes. Positive correlations have been observed between the intronic CNV region of *LEPR* gene and its mRNA levels. In summary, our findings provide a comprehensive appreciation of the full dimension of bovine genetic variations, which may unravel the genetic basis for the improvement of economic phenotypes in cattle.

Supporting information

S1 Fig. Sequencing coverage of the Nanyang and Qinchuan genomes. The x-axis indicated 30 chromosomes (including autosomes and the X chromosome) of the reference genome. The left y-axis represented the length of chromosome (0~160 Mbp), and the right y-axis represented the percentage scale of sequencing coverage (0%~100%). Bars in blue showed the covered region by the sequenced reads, and bars in red showed the uncovered sequence region. The black lines above indicated the percentage of sequencing coverage in each chromosome.
(PDF)

S2 Fig. Novel and common SNPs in each chromosome of Nanyang and Qinchuan cattle genome. A, The number of novel SNPs. B, The number of common SNPs. Blue bars indicated the Nanyang and purple bars showed the Qinchuan genome. Herein, the "novel" means a variant that was not found in dbSNP.
(PDF)

S3 Fig. Distribution of different indel size in whole genome and CDS region. The x-axis indicated indel size of 1 bp, 2 bp, and 3 bp. The left y-axis represented the insertion and deletion in whole genome, and the right y-axis represented the distribution in CDS region.
(PDF)

S4 Fig. Statistics of SNPs and indels in each chromosome. The left y-axis represented the number of genetic variations (0~600,000 for SNPs; 0~10,000 for indels), and the right y-axis represented the scale of chromosome size (0~200 Mbp). Blue and red bars indicated the statistical results of Nanyang and Qinchuan genome, respectively. The green lines represented the length of chromosome.
(PDF)

S5 Fig. Validation of CNVs in individual animals of Nanyang and Qinchuan breeds. The validation results for genic (including gains and losses) and non-genic (including gains and losses) CNV region were provided. The scattergrams of relative copy number were shown for Nanyang (n = 10) and Qinchuan (n = 10). The name of the overlapping genes were given in parentheses for genic CNVs.
(PDF)

S6 Fig. Expression profiling of *LEPR* gene in different tissues of adult cattle. The values are the averages of three independent experiments measured by $2^{-\Delta\Delta Ct}$. Error bars represent the standard deviation (SD) (n = 3), and the relative mRNA expression levels of *LEPR* are normalized to *GAPDH*.
(PDF)

S1 Table. Evaluation of the sequencing data in Nanyang and Qinchuan genome.
(PDF)

S2 Table. The percentage of SNPs with transition, transversion, and heterozygosity in each chromosome of Nanyang and Qinchuan genomes.
(PDF)

S3 Table. The percentage of small indels with insertion, deletion, and both insertion and deletion in each chromosome of Nanyang and Qinchuan genomes.
(PDF)

S4 Table. Statistical number of the genes harboring different mutations.
(PDF)

S5 Table. The list of all CNVs detected in this work. The position of start and end, CNV size, log_2, and p.value were shown for each CNV.
(PDF)

S6 Table. The CNVs in our study were identical or overlapped with those reported in previous papers.
(PDF)

S7 Table. The information of primers used in this study. A total of 20 primer pairs were used for validation in mRNA level, and 2 primer pairs were used for detection in DNA level.
(PDF)

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 30972080, 31272408, 31600617), the Program of the National Beef Cattle Industrial Technology System (CARS-38), the National 863 Program of China (Grant No. 2013AA102505), and the Natural Science Foundation of Jiangsu Province (BK2011206).

Author Contributions

Conceptualization: Yao Xu, Hong Chen.

Formal analysis: Yao Xu, Yu Jiang.

Funding acquisition: Hong Chen.

Project administration: Yao Xu, Tao Shi, Hanfang Cai.

Resources: Xianyong Lan, Xin Zhao, Hong Chen.

Software: Yu Jiang.

Writing – original draft: Yao Xu.

Writing – review & editing: Xin Zhao, Martin Plath, Hong Chen.

References

1. Rutherford SL From genotype to phenotype: buffering mechanisms and the storage of genetic information. Bioessays. 2000; 22: 1095–1105. https://doi.org/10.1002/1521-1878(200012)22:12<1095::AID-BIES7>3.0.CO;2-A PMID: 11084625
2. Boerner V, Johnston D, Wu XL, Bauck S Accuracy of Igenity genomically estimated breeding values for predicting Australian Angus BREEDPLAN traits. J Anim Sci. 2015; 93: 513–521. https://doi.org/10.2527/jas.2014-8357 PMID: 25549982
3. Ellegren H, Sheldon BC Genetic basis of fitness differences in natural populations. Nature. 2008; 452: 169–175. https://doi.org/10.1038/nature06737 PMID: 18337813
4. Hill WG Maintenance of quantitative genetic variation in animal breeding programmes. Livest Prod Sci. 2000; 63: 99–109. https://doi.org/10.1016/S0301-6226(99)00115-3
5. Mackay TF The genetic architecture of quantitative traits. Annu Rev Genet. 2001; 35: 303–339. https://doi.org/10.1146/annurev.genet.35.102401.090633 PMID: 11700286
6. Du Q, Gong C, Wang Q, Zhou D, Yang H, Pan W, et al. Genetic architecture of growth traits in Populus revealed by integrated quantitative trait locus (QTL) analysis and association studies. New Phytol. 2016; 209: 1067–1082. https://doi.org/10.1111/nph.13695 PMID: 26499329
7. Stafuzza NB, Zerlotini A, Lobo FP, Yamagishi ME, Chud TC, Caetano AR, et al. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. PLoS One. 2017; 12. https://doi.org/10.1371/journal.pone.0173954 PMID: 28323836

8. Yang J, Li WR, Lv FH, He SG, Tian SL, Peng WF, et al. Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. Mol Biol Evol. 2016; 33: 2576–2592. https://doi.org/10.1093/molbev/msw129 PMID: 27401233
9. Medeiros de Oliveira Silva R, Bonvino Stafuzza N, de Oliveira Fragomeni B, Miguel Ferreira de Camargo G, Matos Ceacero T, Noely Dos Santos Goncalves Cyrillo J, et al. Genome-Wide Association Study for Carcass Traits in an Experimental Nelore Cattle Population. PLoS One. 2017; 12. https://doi.org/10.1371/journal.pone.0169860 PMID: 28118362
10. Elsik CG, Tellam RL, Worley KC The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 2009; 324: 522–528. https://doi.org/10.1126/science.1169588 PMID: 19390049
11. Eck SH, Benet-Pagès A, Flisikowski K, Meitinger T, Fries R, Strom TM Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biol. 2009; 10: R82. https://doi.org/10.1186/gb-2009-10-8-r82 PMID: 19660108
12. Kawahara-Miki R, Tsuda K, Shiwa Y, Arai-Kichise Y, Matsumoto T, Kanesaki Y, et al. Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. BMC genomics. 2011; 12: 103. https://doi.org/10.1186/1471-2164-12-103 PMID: 21310019
13. Lee K-T, Chung W-H, Lee S-Y, Choi J-W, Kim J, Lim D, et al. Whole-genome resequencing of Hanwoo (Korean cattle) and insight into regions of homozygosity. BMC genomics. 2013; 14: 519. https://doi.org/10.1186/1471-2164-14-519 PMID: 23899338
14. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, et al. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 2012; 22: 778–790. https://doi.org/10.1101/gr.133967.111 PMID: 22300768
15. Stothard P, Choi J-W, Basu U, Sumner-Thomson JM, Meng Y, Liao X, et al. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. BMC genomics. 2011; 12: 559. https://doi.org/10.1186/1471-2164-12-559 PMID: 22085807
16. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9: 356–369. https://doi.org/10.1038/nrg2344 PMID: 18398418
17. Li H, Durbin R Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168
18. Kumar P, Henikoff S, Ng PC Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4: 1073–1081. https://doi.org/10.1038/nprot.2009.86 PMID: 19561590
19. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009; 6: 677–681. https://doi.org/10.1038/nmeth.1363 PMID: 19668202
20. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25: 2865–2871. https://doi.org/10.1093/bioinformatics/btp394 PMID: 19561018
21. Liu GE, Hou YL, Zhu B, Cardone MF, Jiang L, Cellamare A, et al. Analysis of copy number variations among diverse cattle breeds. Genome Res. 2010; 20: 693–703. https://doi.org/10.1101/gr.105403.110 PMID: 20212021
22. Du Z, Zhou X, Ling Y, Zhang Z, Su Z agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010; 38: W64–W70. https://doi.org/10.1093/nar/gkq310 PMID: 20435677
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943
24. Geisler SB, Robinson D, Hauringa M, Raeker MO, Borisov AB, Westfall MV, et al. Obscurin-like 1, OBSL1, is a novel cytoskeletal protein related to obscurin. Genomics. 2007; 89: 521–531. https://doi.org/10.1016/j.ygeno.2006.12.004 PMID: 17289344
25. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res. 2012; 40: D84–D90. https://doi.org/10.1093/nar/gkr991 PMID: 22086963
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25: 25–29. https://doi.org/10.1038/75556 PMID: 10802651
27. Zhou X, Su Z EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC genomics. 2007; 8: 246. https://doi.org/10.1186/1471-2164-8-246 PMID: 17645808

28. Wang J, Jiang J, Fu W, Jiang L, Ding X, Liu J-F, et al. A genome-wide detection of copy number variations using SNP genotyping arrays in swine. BMC genomics. 2012; 13: 273. https://doi.org/10.1186/1471-2164-13-273 PMID: 22726314

29. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006; 444: 444–454. https://doi.org/10.1038/nature05329 PMID: 17122850

30. Chen W-K, Swartz JD, Rush LJ, Alvarez CE Mapping DNA structural variation in dogs. Genome Res. 2009; 19: 500–509. https://doi.org/10.1101/gr.083741.108 PMID: 19015322

31. Ohshima K RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity. Int J Evol Biol. 2013; 2013: 424726. https://doi.org/10.1155/2013/424726 PMID: 23984183

32. Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, et al. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. Genome Biol. 2014; 15: 469. https://doi.org/10.1186/s13059-014-0469-0 PMID: 25239471

33. Liu Y-J, Rocha-Sanchez SM, Liu P-Y, Long J-R, Lu Y, Elze L, et al. Tests of linkage and/or association of the LEPR gene polymorphisms with obesity phenotypes in Caucasian nuclear families. Physiol Genomics. 2004; 17: 101–106. https://doi.org/10.1152/physiolgenomics.00213.2003 PMID: 14970363

34. Xu L, Hou Y, Bickhart DM, Zhou Y, Hay el HA, Song J, et al. Population-genetic properties of differentiated copy number variations in cattle. Sci Rep. 2016; 6. https://doi.org/10.1038/srep23161 PMID: 27005566

35. Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, et al. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Res. 2016; 23: 253–262. https://doi.org/10.1093/dnares/dsw013 PMID: 27085184

36. Keel BN, Lindholm-Perry AK, Snelling WM Evolutionary and Functional Features of Copy Number Variation in the Cattle Genome. Front Genet. 2016; 7. https://doi.org/10.3389/fgene.2016.00207 PMID: 27920798

37. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007; 39: S7–S15. https://doi.org/10.1038/ng2093 PMID: 17597783

38. Singleton AB Exome sequencing: a transformative technology. Lancet Neurol. 2011; 10: 942–946. https://doi.org/10.1016/S1474-4422(11)70196-X PMID: 21939903

39. Crispell J, Zadoks RN, Harris SR, Paterson B, Collins DM, de-Lisle GW, et al. Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. BMC Genomics. 2017; 18: 017–3569. https://doi.org/10.1186/s12864-017-3569-x PMID: 28209138

40. Rubin C-J, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010; 464: 587–591. https://doi.org/10.1038/nature08832 PMID: 20220755

41. Mitchell-Olds T, Willis JH, Goldstein DB Which evolutionary processes influence natural genetic variation for phenotypic traits? Nat Rev Genet. 2007; 8: 845–856. https://doi.org/10.1038/nrg2207 PMID: 17943192

42. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005; 307: 1072–1079. https://doi.org/10.1126/science.1105436 PMID: 15718463

43. Ferraz J, Pinto L, Meirelles F, Eler J, De Rezende F, Oliveira E, et al. Association of single nucleotide polymorphisms with carcass traits in Nellore cattle. Genet Mol Res. 2009; 8: 1360–1366. https://doi.org/10.4238/vol8-4gmr650 PMID: 19937580

44. Matsuhashi T, Maruyama S, Uemoto Y, Kobayashi N, Mannen H, Abe T, et al. Effects of bovine fatty acid synthase, stearoyl-coenzyme A desaturase, sterol regulatory element-binding protein 1, and growth hormone gene polymorphisms on fatty acid composition and carcass traits in Japanese Black cattle. J Anim Sci. 2011; 89: 12–22. https://doi.org/10.2527/jas.2010-3121 PMID: 20852082

45. Islam K, Vinsky M, Crews R, Okine E, Moore S, Crews D, et al. Association analyses of a SNP in the promoter of IGF1 with fat deposition and carcass merit traits in hybrid, Angus and Charolais beef cattle. Anim Genet. 2009; 40: 766–769. https://doi.org/10.1111/j.1365-2052.2009.01912.x PMID: 19466932

46. Pannier L, Mullen A, Hamill R, Stapleton P, Sweeney T Association analysis of single nucleotide polymorphisms in DGAT1, TG and FABP4 genes and intramuscular fat in crossbred Bos taurus cattle. Meat Sci. 2010; 85: 515–518. https://doi.org/10.1016/j.meatsci.2010.02.025 PMID: 20416823

47. Nkrumah J, Li C, Yu J, Hansen C, Keisler D, Moore S Polymorphisms in the bovine leptin promoter associated with serum leptin concentration, growth, feed intake, feeding behavior, and measures of carcass merit. J Anim Sci. 2005; 83: 20–28. https://doi.org/10.2527/2005.83120x PMID: 15583038

48. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 2007; 3: e3. https://doi.org/10.1371/journal.pgen.0030003 PMID: 17206864

49. Uddin M, Thiruvahindrapuram B, Walker S, Wang Z, Hu P, Lamoureux S, et al. A high-resolution copy-number variation resource for clinical and population genetics. Genet Med. 2015; 17: 747–752. https://doi.org/10.1038/gim.2014.178 PMID: 25503493

50. Traherne J Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet. 2008; 35: 179–192. https://doi.org/10.1111/j.1744-313X.2008.00765.x PMID: 18397301

51. Bergen S, O'Dushlaine C, Ripke S, Lee P, Ruderfer D, Akterin S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. Mol Psychiatr. 2012; 17: 880–886. https://doi.org/10.1038/mp.2012.73 PMID: 22688191

52. Larkin DM, Daetwyler HD, Hernandez AG, Wright CL, Hetrick LA, Boucek L, et al. Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. P Natl Acad Sci USA. 2012; 109: 7693–7698. https://doi.org/10.1073/pnas.1114546109 PMID: 22529356

53. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim E-s, et al. Genomic characteristics of cattle copy number variations. BMC genomics. 2011; 12: 127. https://doi.org/10.1186/1471-2164-12-127 PMID: 21345189

54. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 2009; 19: 1527–1541. https://doi.org/10.1101/gr.091868.109 PMID: 19546169

55. Zhang F, Gu W, Hurles ME, Lupski JR Copy Number Variation in Human Health, Disease, and Evolution. Annu Rev Genom Hum G. 2009; 10: 451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217 PMID: 19715442

56. Xu Y, Zhang L, Shi T, Zhou Y, Cai H, Lan X, et al. Copy number variations of MICAL-L2 shaping gene expression contribute to different phenotypes of cattle. Mamm Genome. 2013; 24: 508–516. https://doi.org/10.1007/s00335-013-9483-x PMID: 24196410

57. Xu Y, Shi T, Cai H, Zhou Y, Lan X, Zhang C, et al. Associations of MYH3 gene copy number variations with transcriptional expression and growth traits in Chinese cattle. Gene. 2014; 535: 106–111. https://doi.org/10.1016/j.gene.2013.11.057 PMID: 24316128

58. Liu G, Van Tassell C, Sonstegard T, Li R, Alexander L, Keele J, et al. Detection of germline and somatic copy number variations in cattle. Dev Biol. 2008; 132: 231–237. https://doi.org/10.1159/000317165

59. Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon D, Vieaud A, et al. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet. 2009; 5: e1000512. https://doi.org/10.1371/journal.pgen.1000512 PMID: 19521496

60. Friedman JM The function of leptin in nutrition, weight, and physiology. Nutr Rev. 2002; 60: S1–S14. https://doi.org/10.1301/002966402320634878

61. Jeon J-P, Shim S-M, Nam H-Y, Ryu G-M, Hong E-J, Kim H-L, et al. Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. BMC genomics. 2010; 11: 426. https://doi.org/10.1186/1471-2164-11-426 PMID: 20624279

62. Hu Z-L, Fritz ER, Reecy JM AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. Nucleic Acids Res. 2007; 35: D604–D609. https://doi.org/10.1093/nar/gkl946 PMID: 17135205