Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Diversity of copy number variation in a worldwide population of sheep

Liu Yang[a,b], Lingyang Xu[b,c], Yang Zhou[c,d], Mei Liu[c,d], Lei Wang[e], James W. Kijas[f], Hongping Zhang[a], Li Li[a,*], George E. Liu[c,**]

[a] Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu 611130, China
[b] Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, China
[c] Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705, USA
[d] College of Animal Science and Technology, Northwest A & F University, Shaanxi Key Laboratory of Agricultural Molecular Biology, Yangling, Shaanxi 712100, China
[e] CapitalBio Technology, Beijing 101111, China
[f] CSIRO Agriculture, Brisbane 4065, QLD, Australia

ARTICLE INFO

Keywords:
Copy number variation
Diversity
Evolution
Selection
Sheep

ABSTRACT

Copy number variation (CNV) represents a major source of genomic variation. We investigated the diversity of CNV distribution using SNP array data collected from a comprehensive collection of geographically dispersed sheep breeds. We identified 24,558 putative CNVs, which can be merged into 619 CNV regions, spanning 197 Mb of total length and corresponding to ~6.9% of the sheep genome. Our results reveal a population differentiation in CNV between different geographical areas, including Africa, America, Asia, Southwestern Asia, Central Europe, Northern Europe and Southwestern Europe. We observed clear distinctions in CNV prevalence between diverse groups, possibly reflecting the population history of different sheep breeds. We sought to determine the gene content of CNV, and found several important CNV-overlapping genes (*BTG3*, *PTGS1* and *PSPH*) which were involved in fetal muscle development, prostaglandin (PG) synthesis, and bone color. Our study generates a comprehensive CNV map, which may contribute to genome annotation in sheep.

1. Background

Sheep (*Ovis aries*) were domesticated in the Fertile Crescent approximately 9000 years ago [1,2], and have since become an important farm animal. The phenotypic and genetic variability that exists between sheep breeds is due to adaptation and artificial selection for animal production including meat, wool and milk [3]. Recently, high-throughput single nucleotide polymorphism (SNP) arrays have facilitated population genetics studies to improve our understanding of the genetic mechanism underlying complex economic and adaptive traits in domesticated animals [4,5]. Notably, the availability of dense SNP datasets has dramatically improved our understandings about genetic diversity, population history admixture, selection signatures, and other features in local and worldwide sheep populations [6–11].

On the other hand, our knowledge of genome function and evolution is still limited to SNP which, although widely used in genome research in farm animals, is just one type of common genomic variations. Copy number variation (CNV) is widely dispersed in mammalian genomes and also independently contributes to phenotypic diversity and disease susceptibility [12–14]. Recently, many studies have proposed that CNV can be used to study population genetics and show lineage-specific selection signatures in human [15–18], zebrafish [19], stickleback [20], and cattle [21]. However, for sheep, previous studies mostly focused on CNV discovery, which relied on using comparative genomic hybridization (CGH) and SNP arrays [22–25]. Thus, the features of sheep CNV at the population level are not well understood. Additionally, besides SNP, CNV is a separated type of important genomic variants. Due to CNV's less known linkage disequilibrium (LD) patterns, CNV-based population genomics results could offer additional new insights for functional and evolutionary studies in sheep. Therefore, the investigation of population genetics based on CNV could facilitate our understanding of evolution and selection aspects of the sheep genome.

In this study, we used the Sheep HapMap dataset to investigate CNV in the worldwide sheep populations. We utilized PennCNV to detect CNV using Illumina Ovine SNP50 genotyping data, and performed CNV-based population genetics and selection sweep analysis. With a large-scale sheep CNV map, our results provide comprehensive CNV information in sheep, and provide potential candidate variations for further exploration on the roles of CNV underlying important traits and

* Corresponding author.
** Corresponding author at: Animal Genomics and Improvement Laboratory, USDA-ARS, Building 306, Room 111, BARC-East, Beltsville, MD 20705, USA.
E-mail addresses: zhp@sicau.edu.cn (H. Zhang), lily@sicau.edu.cn (L. Li), George.Liu@ars.usda.gov (G.E. Liu).

http://dx.doi.org/10.1016/j.ygeno.2017.09.005
Received 31 March 2017; Received in revised form 24 August 2017; Accepted 11 September 2017

Table 1
Summary of CNV and CNVR identified in 2111 samples derived from the sheep HapMap populations. Samples were divided into African (AFR), American (AME), Asian (ASI), Central European (CEU), Northern European (NEU), Southwestern Asian (SAS), and Southwestern European (SEU) groups. CNVR in this table represent non-redundant CNVR counts after merging both gain and loss CNVs identified within each group.

Groups	Sample size	CNV						CNVR			
		Count/ average[a]	Gain/ average	Loss/ average	Length (bp)	Average	SD[b]	Count[c]	Length (bp)	Average	SD[b]
African	117	1470/12.56	61/0.52	1409/12.04	216,493,684	147,274.61	143,971.56	212	53,261,080	251,231.51	285,165.74
American	213	2514/11.80	155/0.73	2359/11.08	346,200,939	137,709.20	115,374.43	211	49,087,317	232,641.31	250,620.72
Asian	206	2574/12.50	152/0.74	2422/11.76	372,547,204	144,734.73	118,862.63	233	55,001,353	236,057.31	253,896.20
Central European	242	2856/11.80	141/0.58	2715/11.22	386,881,910	135,462.85	91,545.60	170	32,014,638	188,321.40	136,481.48
Northern European	496	5763/11.62	245/0.49	5518/11.13	824,741,758	143,109.80	119,099.42	334	86,524,749	259,056.13	319,535.30
Southwestern Asian	136	1953/14.36	73/0.54	1880/13.82	332,389,280	170,194.20	161,479.07	255	70,938,258	278,189.25	265,099.73
Southwestern European	701	7428/10.60	526/0.75	6902/9.85	977,647,582	131,616.53	98,442.08	343	81,019,492	236,208.43	249,716.33

The numbers of Average are divided or normalized by sample counts except that the lengths are average lengths normalized by CNV and CNVR counts.
[a] At sample level, each sample has 12.56 (1470/117) CNVs, 0.52 (61/117) and 12.04 (1409/117) averagely for African.
[b] Standard deviation.
[c] These numbers are non-redundant CNVR counts.

evolutionary adaptation in sheep.

2. Method and materials

2.1. Selecting populations and animals

Data from sixty-eight breeds and 2254 animals are retrieved from the Sheep HapMap dataset which has been previously published [6]. This included 136 animals from African (AFR) breeds, 222 animals from American (AME) breeds, 211 animals from Asian (ASI) breeds, 198 animals from Southwestern Asian (SAS) breeds, 242 animals from Central European (CEU) breeds, 533 animals from Northern European (NEU) breeds, and 712 animals from Southwestern European (SEU) breeds. It is worth noting that for many breeds, individuals were sampled from more than one continent to explore within-breed genetic diversity. The number of animals per population and geographic origin of breed development were described in [6]. All chosen samples had a genotyping success rate of > 99%.

2.2. CNV detection with PennCNV

Signal intensity ratios (log R Ratio: LRR) and allelic frequencies (B allele frequency: BAF) were retrieved using Illumina GenomeStudio1.0 software for each SNP. The population frequency of B allele (PFB) file was calculated based on the BAF of each marker in each population. The sheep GC model file was generated using cal_gc_snp.pl with default settings (http://penncnv.openbioinformatics.org/en/latest/misc/faq/). CNVs were inferred within each individual using PennCNV software based on OARv1.0 [26]. PennCNV quality filters were subsequently applied as follows: We used high quality samples with a standard deviation (SD) of LRR < 0.35 and with the parameter set: BAF drift as 0.01 and waviness factor value between − 0.05 and 0.05, respectively. Appropriate LRR adjustments based on the GC model were incorporated in PennCNV. In addition, we used the program argument: the “lastchr 26” in the “detect” argument for specific CNVs. CNV regions (CNVR) were determined by aggregating overlapping CNVs identified in different animals, as reported previously [27]. For construction of the CNVR map, we classified the status of these CNVR into three categories, ‘Loss’ (CNVR containing deletion), ‘Gain’ (CNVR containing duplication) and ‘Both’ (CNVR containing both deletion and duplication).

2.3. Gene annotation and PANTHER analysis

We retrieved and annotated RefSeq gene overlapping CNV regions, using OARv1.0 from the UCSC Genome Browser. We performed gene ontology (GO) enrichment analysis using PANTHER with a genome-wide gene list based on the cattle (*Bos taurus*) as sheep's list was not present. We only considered terms with gene count > 5 and adjusted *p*-value < 0.05 (Bonferroni correction for multiple testing).

2.4. Population differential analysis

To explore lineage-specific CNVs, we first divided the Sheep HapMap samples into seven groups according to the geography distribution. They included African (AFR), American (AME), Asian (ASI), Southwestern Asian (SAS), Central European (CEU), Northern European (NEU), and Southwestern European (SEU) groups. We then constructed a comparative CNVR map across these seven groups. The frequency of a CNV within each CNVR was measured and utilized as the CNV characteristics for comparison among seven groups. To explore the potential differences involved with selection pressure for CNVs, we estimated the CNV frequency per group and the variance across all seven groups. Based on the frequency across seven groups, Euclidian distances were calculated. Using Ward's method as the linkage criteria, hierarchical cluster analysis was performed using 19 or 39 CNVR at top 5% or top 10% of the variances of frequency, respectively.

3. Results

3.1. Identification of copy number variation in worldwide sheep populations

A total of 52,094 autosomes markers were selected from the OvineSNP50 assay for CNV analysis. Using PennCNV, we performed large-scale CNV explorations in 2254 sheep derived from the Sheep HapMap Project dataset. After quality filtering, we identified 24,588 CNVs across 2111 individuals from 68 breeds with total length 3457 Mb and average length for each individual was 1.6 Mb. Among them, we observed 1470 CNVs in AFR, 2514 in AME, 2574 in ASI, 1953 in SAS, 2856 in CEU, 5763 in NEU, 7428 in SEU groups (see Table 1 for summary statistics of CNV and CNVR).

To explore the geographic pattern of CNV across the seven groups, we constructed violin plots using the CNV length and the number of SNP within CNV events. This identified patterns relating to CNV length and SNP count across groups (Fig. 1). CNV length generally showed slight difference, while the Central European (CEU) group was notably less variable than the Northern European (NEU) group (Fig. 1A) and the total numbers of SNP covered by each CNV per individual shows similar patterns to those observed for CNV lengths (Fig. 1B).

To compare the frequency of CNV across different groups, we first merged CNVs for each individual into non-redundant CNVR within