Genomics 106 (2015) 286–294

Contents lists available at ScienceDirect

Genomics

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

Resequencing diverse Chinese indigenous breeds to enrich the map of genomic variations in swine

Huimin Kang [a,1], Haifei Wang [a,1], Ziyao Fan [a,1], Pengju Zhao [a], Amjad Khan [a], Zongjun Yin [b], Jiafu Wang [c], Wenbin Bao [d], Aiguo Wang [a], Qin Zhang [a], Jian-Feng Liu [a,*]

[a] National Engineering Laboratory for Animal Breeding, Key Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China
[b] College of Animal Science and Technology, Anhui Agricultural University, Hefei 230036, China
[c] School of Animal Science, Guizhou University, Guiyang 550025, China
[d] College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, China

ARTICLE INFO

Article history:
Received 27 September 2014
Received in revised form 10 August 2015
Accepted 11 August 2015
Available online 18 August 2015

Keywords:
SNV
INDEL
NGS
Pigs

ABSTRACT

To enrich the map of genomic variations in swine, we randomly sequenced 13 domestic and wild individuals from China and Europe. We detected approximately 28.1 million single nucleotide variants (SNVs) and 3.6 million short insertions and deletions (INDELs), of which 2,530,248 SNVs and 3,456,626 INDELs were firstly identified compared with dbSNP 143. Moreover, 208,687 SNVs and 24,161 INDELs were uniquely observed in Chinese pigs, potentially accounting for phenotypic differences between Chinese and European pigs. Furthermore, significantly high correlation between SNV and INDEL was witnessed, which indicated that these two distinct variants may share similar etiologies. We also predicted loss of function genes and found that they were under weaker evolutionary constraints. This study gives interesting insights into the genomic features of the Chinese pig breeds. These data would be useful in the establishment of high-density SNP map and would lay a foundation for facilitating pig functional genomics study.

1. Introduction

With the maturation of next-generation sequencing (NGS) technology, whole genome sequencing offers a feasible way to unleash the potential of genomics, greatly facilitating identification of genomic variations. The availability of draft genome sequence and strategies for massive data analyses make it possible to identify various genomic variations efficiently and accurately, such as single nucleotide polymorphisms (SNPs), short insertions and deletions (INDELs), copy number variations and other types of structure variations. Especially, a great success has been achieved in identification of a huge number of SNPs and INDELs based on NGS data in humans [1] as well as livestock [2–4]. Plenty of these genetic variations were elucidated to be functionally associated with common diseases in humans [5,6] and economically important traits in livestock [7]. Identification of genetic variations over the entire genome is the fundamental task for causal variant discovery and related studies.

Pig (*Sus scrofa [S. scrofa]*) was domesticated about 10,000 years ago and has undergone strong selection under different environments, resulting in various phenotypic features and genetic adaptations. Pigs have been considered as biomedical models for various human diseases [8] as well as one of the most economically important livestock. To get insights into the genetics associated with phenotypic variation and potential applications in pig genetics and breeding, it is imperative to appreciate genetic variations across various populations/breeds [9]. This can also provide an important resource for comparative investigations in human genetics [10].

In pig genome, massive parallel sequencing has been applied to the whole genome SNP discovery [3,11–13]. Furthermore, release of the second high quality assembly (Sscrofa10.2) provided a vital source for further improvements of this important domestic animal [13]. About 51.2 million single nucleotide variants (SNVs) have been identified across the genome so far, which are stored in the Database of Short

Abbreviations: SNVs, single nucleotide variants; INDELs, short insertions & deletions; NGS, next-generation sequencing; SNPs, single nucleotide polymorphisms; *Sus scrofa, S. scrofa*; dbSNP, Database of Short Genetic Variation; GWAS, genome-wide association study; GS, genomic selection; QUAL, quality score; GO, Gene Ontology; CDS, coding sequences; UTRs, untranslated regions; ncRNA, non-coding RNA; refGene, RefSeq Genes; OMIM, Online Mendelian Inheritance in Man; OMIA, Online Mendelian Inheritance in Animals; FPR, false positive rate; LoF, loss of function.

* Corresponding author at: China Agricultural University (West District), College of Animal Science and Technology, Room 455, 2 Yuanmingyuan West Road, Beijing 100193, China.

E-mail addresses: nongdaxiaokang@126.com (H. Kang), wanghaiffei@126.com (H. Wang), nuclearfan@cau.edu.cn (Z. Fan), zhaopengju2014@gmail.com (P. Zhao), amjadkhanqau@gmail.com (A. Khan), yinzongjun@ahau.edu.cn (Z. Yin), jfwang@gzu.edu.cn (J. Wang), wbbbao@yzu.edu.cn (W. Bao), agwang@cau.edu.cn (A. Wang), qzhang@cau.edu.cn (Q. Zhang), liujf@cau.edu.cn (J.-F. Liu).

[1] Equal contributors.

http://dx.doi.org/10.1016/j.ygeno.2015.08.002
0888-7543/© 2015 Elsevier Inc. All rights reserved.

Genetic Variation (dbSNP) build 143 (http://www.ncbi.nlm.nih.gov/snp/). Up to now, whole-genome resequencing has been widely used to explore genomic variations in pigs from several main domestic breeds and wild boar populations [3,11–14].

A total of 158 Chinese indigenous pig breeds are documented in the Domestic Animal Diversity Information System of the Food and Agriculture Organization (http://dad.fao.org/). These breeds are traditionally divided into six types according to the genetic similarity and geographical origin, i.e., South China Type, Lower Changjiang River Basin Type, Central China Type, North China Type, Southwest Type, and Plateau Type [15]. High genetic diversity exists in Chinese pig breeds [16], providing abundant and invaluable resource for pig genomic studies. Since the structure and sequence across different populations/breeds are highly variable, generation of whole genome sequence from a number of individuals with different geographical and genetic backgrounds would be necessary for a more comprehensive understanding of genomic variations. However, the aforementioned six types of Chinese indigenous breeds have not been exhaustively explored for genomic variations.

Considering the case that pigs of Asian origin, mainly the Chinese indigenous pigs, and European pigs are two main clades of domestic pigs [17], sequencing genomes of six Chinese indigenous types of pig populations will greatly enrich their genomic variation map, and would further enhance our understanding of their genomic features and phenotypic variations.

Towards this end, we have conducted the current study focusing on exploiting genome-wide SNVs/INDELs across nine individuals from six types of Chinese indigenous populations, one southern Chinese wild boar and three pigs of European origin based on NGS technology. These results would contribute to a more comprehensive dbSNP of pig and offer an updated genomic resource, and would further advance comparative genomic studies as well as genome-wide association study (GWAS) and genomic selection (GS) in pig breeding.

2. Materials and methods

2.1. Ethics statements

The entire procedure for collection of the ear tissue samples of all animals was performed in strict accordance with the protocol approved by the Institutional Animal Care and Use Committee of China Agricultural University.

2.2. Animals

Collectively we sampled 13 unrelated pigs originating from ten distinct breeds/populations for sequencing. These animals comprised one southern Chinese wild boar, three modern commercial pigs (one Landrace, one Duroc and one Yorkshire), and nine Chinese indigenous pigs. The Duroc, Yorkshire and Landrace individuals were representatives of modern commercial breeds. These nine Chinese indigenous pigs, as representatives of all the six Chinese distinct domestic population types, were comprised of Tibetan (Plateau Type, $n = 2$), Diannan small-ear (South China Type, $n = 2$), Meishan (Lower Changjiang River Basin Type, $n = 2$), Min (North China Type, $n = 1$), Daweizi (Central China Type, $n = 1$), and Rongchang (Southwest Type, $n = 1$). All of the 13 experimental animals were females, except the Rongchang pig and one of the Diannan small-ear pigs. The features of these six types of Chinese indigenous populations were illustrated elsewhere [18].

2.3. Genomic DNA preparation and sequencing

Genomic DNA of 13 individuals was extracted from ear tissue using the Qiagen DNeasy Tissue Kit (Qiagen, Hilden, Germany) according to the recommended protocols. The isolated DNA was analyzed with agarose gel electrophoresis and a NanoDropTMND-2000c Spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Subsequently, genomic DNA of each individual was randomly fragmented and DNA fragments of the desired length were gel purified. Adapter ligation and DNA cluster preparation were performed and the fragments were sequenced using the Illumina HiSeq 2000 sequencing system.

We prepared double libraries of genomic DNA for paired-end sequencing of each individual to ensure a high level of coverage. Quality filtering of the raw data was performed by following two procedures: firstly, reads that had been polluted by adapter sequences were eliminated, and then reads containing more than 50% low quality bases (quality value ≤ 5) or more than 10% N bases were removed.

2.4. SNV and INDEL identification

The generated paired-end reads were firstly aligned to Sscrofa10.2 downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) using a fast and accurate short read alignment program BWA [19]. Three modules involved in BWA were run sequentially for read alignment, i.e., commands index with option *bwtsw*, aln and sampe with default options. Then we ran Picard (http://picard.sourceforge.net) to remove duplicate reads and GATK [20] to realign the remaining reads locally. Finally for all the sequenced genomes, we called SNVs and INDELs through SAMtools [21] and BCFtools (distributed with SAMtools). The procedure for calling SNVs and INDELs are listed as follows.

In the SAMtools running with command mpileup, we set the corresponding options of disabling probabilistic realignment for computation of base alignment quality (−B), downgrading mapping quality of reads containing excessive mismatches (−C 50) and a minimum mapping quality of 20 for an alignment to be used (−q 20).

In the step of running BCFtools and *vcfutils.pl* (distributed with SAMtools), we set the following parameters for SNV and INDEL calling and initial filtering: the minimum depth was set to four for autosomes and X chromosome in female (−d 4), and two for sex chromosomes in male; the minimum root mean square mapping quality for SNVs was 20 (−Q 20); multiple INDELs occurring within a 20 bp window were filtered out (−W 20) and SNVs within 3 bp around an INDEL (−w 3) were also removed. Additionally, the cut-off was set to about 2.5–3 times the average depth in the SNV and INDEL calling. As average depth varied with the individuals sequenced, we set different cut-off according to the average depth of each individual. Specifically, for autosomes and X chromosome in female, the cut-off was set to 40 (−D 40) for the southern Chinese wild boar and the Rongchang pig, while it was 30 (−D 30) for other individuals. For sex chromosomes in males, the cut-off was half of that for autosomes of the same genome.

After the above two steps of filtration, we ran a self-developed Perl script to remove variants with overall quality score (QUAL) below 40 or genotype quality less than 20, and then employed GATK to discard those three or more SNVs which occurred within a 10 bp length window of the chromosomal region. Heterozygous SNVs and INDELs found in non-pseudoautosomal region on X chromosome in males were also removed from our study.

We also downloaded the publicly available whole-genome sequence data of 14 Yorkshire and five Landrace [13] (Table S1) (http://www.ncbi.nlm.nih.gov/sra/), which was combined with three western genomes sequenced in our study to filter for variants specific to Chinese pigs. Raw data was assessed with the NGS QC Toolkit [22] to remove reads containing more than 30% low quality bases (quality value ≤5), trim bases having PHRED quality score less than 5 at 3′ end of the read, and discard reads shorter than 70 bp. Reads were aligned with BWA as described above. Then we employed Picard to remove duplicates and GATK to realign reads locally and recalibrate base quality score. Population-based variants were called with SAMtools. With SAMtools mpileup, we set −B, −C 50 and −q 20 as described above and −d 10,000 (reading maximally 10,000 reads per input BAM file). In the step of filtration with *vcfutils.pl*, we set the same parameters as