564. A pan-genome of commercial pig breeds

M.F.L. Derks[1,2*], A. Boshove[2], B. Harlizius[1], E. Sell-Kubiak[5], M.S. Lopes[1,4], E. Grindflek[3], E. Knol[1], M.A.M Groenen[2] and A.B. Gjuvsland[3]

[1]Topigs Norsvin Research Center, Schoenaker 6, 6641 SZ Beuningen, the Netherlands; [2]Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [3]Norsvin SA, Hamar, Norway; [4]Topigs Norsvin, Curitiba, Paraná, Brazil; [5]Poznań University of Life Sciences, Department of Genetics and Animal Breeding, Wołyńska Poznań, Poland; martijn.derks@wur.nl

Abstract

Genomics related research in animal breeding is usually performed by comparing genomic information to an existing reference genome. However, even if the reference genome is of high quality, the use of a single reference genome has clear drawbacks. Therefore, the breeding community is shifting towards the construction of a pan-genome for important agricultural species. In this study we produced a pig pan-genome based on four different breeds (Landrace, Large White, Synthetic, Duroc) using the nanopore long read sequencing technology. We produced chromosome arm level assemblies comparable to the current *Sus scrofa* 11.1 reference genome. We identified between breed structural variation, which gives a unique insight in the genomic structural variation that define and differentiates breeds. The pig pan-genome will facilitate the discovery of novel variation providing a unique fundamental insight into breed genomic characteristics, which can subsequently be utilized for breeding.

Introduction

In recent decades, high quality reference genomes have become available for most important livestock species. The availability of the pig reference genome (of Duroc origin) together with gene annotation have revolutionized pig genomics and genetics research over the past decade (Groenen 2016). Despite the high quality of the reference genome, working with a single reference genome also has clear drawbacks (Ballouz *et al.* 2019), meaning that sequences deviating considerably from the reference will be interpreted as low-quality, so-called reference bias. One consequence of much (structural) variation is often missed. Structural variation includes various types of variation in which longer stretches of DNA are altered. Structural variants can have a large effect on phenotypes but they are often ignored or remain unidentified (Bickhart and Liu 2014). The developments in long-read sequencing technology, now enables these shortcomings to be addressed. Hence, current focus shifts towards the generation of a pangenome sequence (i.e. all genes and genetic variation within a species that is built from the alignment of different reference genomes) in many important agricultural species, including pigs (Tian *et al.* 2020). The pig pan-genome is particularly useful to identify presence/absence gene variations, structural variation, and other miscellaneous variations. In this study we aim to construct breed specific reference genomes of four commercial pig breeds and assess their structural variation.

Materials & methods

Samples. DNA from 30 individuals was sequenced using the nanopore technology to obtain long reads. Four individuals were sequenced on three flowcells to obtain high coverage sequence data to build breed specific reference genomes (Figure 1). The breeds comprise two dam lines (Large White and Landrace) and two sire lines (Duroc and Synthetic). Additionaly 26 animals were sequenced using one flowcell each. The Circulomics kit was used for DNA extraction. All flow cells were subject to three loadings to obtain optimum output. The average number of gigabases sequenced was 120 Gb, reflecting an average coverage of 48X per flowcell. The average read N50 was 42 kb.

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_564, © M.F.L. Derks *et al.* 2022

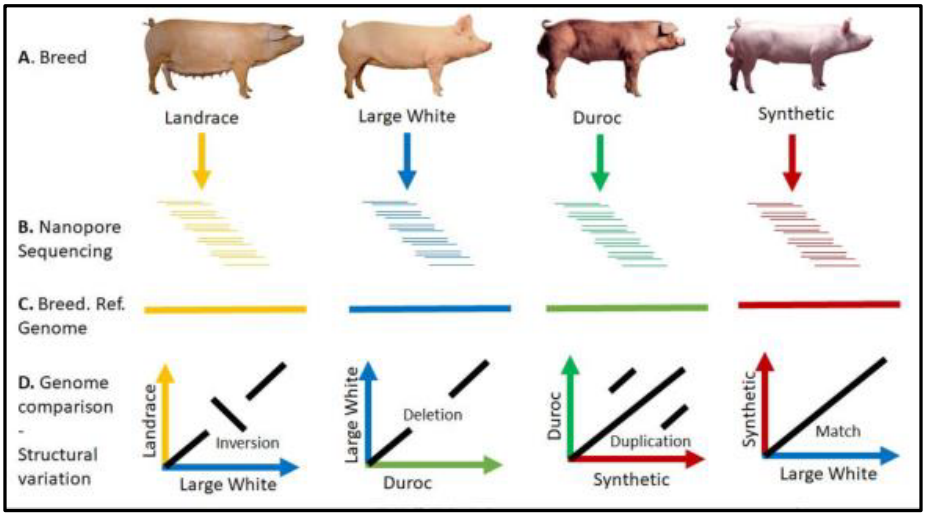

Figure 1. (A) Different pig breeds used in commercial pig breeding; (B) Long read sequencing; (C) Assembly of the long reads to produce breed specific reference genomes; (D) Genome comparison and types of structural variation between the breeds.

Genome Assembly and completeness. We used Flye 2.9 to produce a de novo assembly with the nanopore long reads to construct four breed-specific reference genomes (Kolmogorov *et al.* 2019). Further scaffolding was performed using ntLink (Coombe *et al.* 2021). Polishing was performed using Polca with Illumina short read data from the same individuals (Zimin and Salzberg 2020). The assemblies were anchored to the *Sus scrofa* 11.1 reference genome using RagTag (Alonge *et al.* 2019). The annotation from *Sus scrofa* 11.1 was lifted to the breed-specific reference genomes using Liftoff (Shumate and Salzberg 2020). We assessed the completeness using the BUSCO pipeline (Seppey *et al.* 2019) with the mammalian dataset.

Structural variation and functional prediction. We ran Syri to assess (structural) variation between the line specific genomes (including the *Sus scrofa* 11.1 reference genome) in a pairwise setting (Goel *et al.* 2019). Syri provides VCF files with all identified (structural) variation. We annotated the structural variation for their potential functional consequences using the Ensembl Variant Effect Predictor (VEP) build 104 (McLaren *et al.* 2016). For the one flowcell samples we mapped the samples to the breed reference genome using minimap2 (Li 2018). Subsequently we used Sniffles to assess structural variation between the sample and the breed specific reference genome.

Results

Breed-specific reference genomes. We constructed four breed specific reference genomes. We were able to generate chromosome arm level assemblies for each breed. The acrocentric chromosomes were covered in single scaffolds, whereas the other chromosomes were comprised of 2-4 scaffolds per chromosome. The N50 ranged from 66 to 84 Mb and the contig N50 ranged from 36 to 44 Mb, comparable to the contig N50 of the Sus scrofa 11.1 reference genome. The assembly statistics are given in Table 1.

The assembly completeness was 96.4% for all four breed specific reference genomes using the mammalian BUSCO dataset. The assembly completeness improved significantly after polishing with the Illumina short reads (from ~92 to ~96%).

Table 1. Assembly statistics.

	Sscrofa11.1	Duroc	Landrace	Large White	Synthetic
Assembly length	2,501,921,388	2,478,252,964	2,457,756,321	2,467,972,811	2,476,650,521
Longest sequence	274,330,532	159,819,763	160,007,938	131,827,189	173,409,986
# Scaffolds	613	503	503	607	519
Scaffold N50	138,966,237	84,174,150	77,125,246	66,289,209	79,784,715
Scaffold L50	7	11	13	15	10
# Contigs	1,124	689	683	1,005	694
Contig N50	48,231,277	43,213,177	41,185,454	36,125,157	44,463,933
Contig L50	14	18	20	24	17

Structural variation compared to the Sus scrofa 11.1 reference genome. We used Syri to assess structural differences between the breed specific reference genomes and the 11.1 reference genome (Table 2). The Duroc genome has the least differences with the reference genome especially for the SNPs, inversions, and deletions category. Landrace has a higher number of deletions compared to the other breed-specific reference genomes. This is because we don't have short read data for Landrace (yet) and therefore this breed-specific reference genome remains unpolished, reflected in the number of indels (insertions, deletions).

Large structural variations. We identified various large structural differences between the breed specific reference genomes. First, we identified a 2.8 Mb inversion on chromosome 6 present in all four breed specific reference genomes, but not present in *Sus scrofa* 11.1. This is likely an assembly error in the reference genome. Next we identified a Large White specific 6.2 Mb inversion on chromosome 17. In addition, the start of chromosome 10 is incomplete in the 11.1 reference genome. This first 2 Mb of this chromosome is present in an unplaced scaffold in 11.1 but it's attached to chromosome 10 in our four breed specific reference genomes.

Table 2. Structural variation between the breed specific genomes and the Sus scrofa 11.1 reference genome.

Variation	Large White	Duroc	Synthetic	Landrace
Inversions	147	105	129	139
Translocations	37	55	75	52
Duplications (ref)	31	3	15	15
Duplications (qry)	323	457	477	433
SNPs	7 M	5 M	6.8 M	6.8 M
Insertions	1.5 M	1.3 M	1.3 M	1.1 M
Deletions	785 K	667 K	812 K	4.2 M
Copygains	118	90	113	123
Copylosses	126	90	101	107

Figure 2. Large white specific (Synthetic is Large White based) inversion on chromosome 17.

Interestingly, we identified a translocation of the complete KIT locus (length 490 kb). The locus is translocated to position 55,035,942 on chromosome 8 (from 41,294,443). That translocation was not identified in the Duroc line specific genome.

Genes affected by structural variants. Table 3 shows the set of structural variants, the type of variants and the genes affected by the variants. Last four columns show the presence of the structural variants in breeds.

Discussion

The power of a commercial pig pangenome. This pig pangenome is a useful resource to identify structural variation between breeds as well as within breeds. Moreover, mapping of short read data will be improved once it is mapped to the pangenome graph instead of a single reference genome. Note that we cannot exclude that some of differences stem from assembly errors despite our high quality chromosome arm assemblies.

Genes affected by structural variation. We identified a set of genes affected by structural variation. Most of the variants disrupt part of the gene. It still remains to be investigated what the exact consequence of the variant is. It is possible that the variant only affects a single isoform of the affected gene. Nevertheless several shared variants are found between different breeds. These variants could have originated before the breed differentiation.

Within breed structural variation. Assessing within breed structural variation that segregates at moderate or low frequency will be highly valuable for breeding purposes. After identification, we plan to assess the phenotypic consequence of the structural variants and potentially add the variants to a SNPchip

Table 3. List of genes affected by structural variation (coding sequence). Notal means not aligned, and cpl means copy loss in query.

Chr	SV	Length	Gene symbol	LW	Du	Sy	LR
12	DEL	~8.8KB	ACSF2	**X**	**X**	**X**	**X**
2	DEL	~19.7KB	ARHGAP26	X		X	X
12	NOTAL	~200KB	CA10	X			X
6	DUP	~4.2KB	CCDC30		X	X	X
14	CPL	~27KB	GPAM		X		X
X	DUP	~3.7KB	GPM6B		X	X	X
8	NOTAL	~75KB	KCNIP4	**X**	**X**	**X**	**X**
16	NOTAL	~93KB	LIFR		X	X	
4	DEL	~4.6KB	MAN1A2	X		X	
8	NOTAL	~20KB	PIGG	**X**	**X**	**X**	**X**
2	DEL	~8.5KB	SAFB2	X	X		
3	DEL	~10KB	SEPTIN12	**X**	**X**	**X**	**X**
7	DEL	~6KB	TNXB		X	X	
5	DEL	~900bp	TP23		X	X	
9	DEL	~4KB	TRIM77		X	X	
14	DEL	~24KB	USP30		X	X	
X	DUP	~1.3KB	ZFX		X	X	X
6	DEL	~19KB	ZNF543	**X**	**X**	**X**	**X**

used for routine genotyping and enable selecting on identified variants. Moreover, we now map sequence reads to a reference genome constructed from an animal that belongs to the same population. This will greatly improve structural variation discovery because there are less structural differences between the sample and the reference genome.

References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* 20: 224.

Ballouz S, Dobin A, Gillis JA. 2019. Is it time to change the reference genome? *Genome Biol* 20: 159.

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. *Front Genet* 5: 37.

Coombe L, Li JX, Lo T, Wong J, Nikolic V, Warren RL, Birol I. 2021. LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* 22: 534.

Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* 20: 277.

Groenen MA. 2016. A decade of pig genome sequencing: a window on pig domestication and evolution. *Genet Sel Evol* 48: 23.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37: 540-546.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094-3100.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* 17: 122.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* 1962: 227-245.

Shumate A, Salzberg SL. 2020. Liftoff: accurate mapping of gene annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1016.

Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y *et al.* 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci China Life Sci* 63: 750-763.

Zimin AV, Salzberg SL. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol* 16: e1007981.