562. Detection and characterisation of *de novo* structural variants in pigs

M. Steensma[1*], Y.L. Lee[1], A.C. Bouwman[1], C. Pita Barros[1], M.F.L. Derks[1,2], M.C.A.M. Bink[3], B. Harlizius[2], A.E. Huisman[3], R.P.M.A. Crooijmans[1], M.A.M. Groenen[1], H.A. Mulder[1] and C.M. Rochus[4]

[1]Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2]Topigs Norsvin Research Center, Schoenaker 6, 6641 SZ Beuningen, the Netherlands; [3]Hendrix Genetics, P.O. Box 114, 5830 AC Boxmeer, the Netherlands; [4]University of Guelph, Centre for Genetic Improvement of Livestock, 50 Stone Rd E, N1G 2W1, Guelph, ON, Canada; marije.steensma@wur.nl

Abstract

De novo mutations arising in the germline add to genetic variation. The number of *de novo* mutations occurring every generation, especially structural variants, has not been well studied in most species, including livestock. We used whole-genome sequencing from 46 pig trios from two commercial lines to identify *de novo* structural variants (dnSVs) present in the offspring. We characterised these dnSV by identifying their parent-of-origin, predicting their causal mechanisms, and identifying their functional annotations. We identified four dnSVs, including two clusters of mutations. One of these clusters contained a deletion, and three duplications, one of which was inverted. This cluster was the only dnSV that could be phased and was located in the paternal haplotype of the proband. All four identified dnSVs were located within the introns of genes. Our study is the first of its kind to identify and characterise dnSVs using whole genome shotgun sequence data in pigs.

Introduction

Mutation is the fundamental source of genetic variation. There are two types of mutations that can be found in the germline, and therefore can be transmitted to the next generation: *de novo* mutations (DNM) which arise in the germline during gametogenesis (Bishop *et al.*, 2020) and mosaic mutations which have arisen early in the zygote cell during embryogenesis (Jin *et al.*, 2018; Sasani *et al.*, 2019). The genetic variation introduced by mutations, allows for continued selection and adaption in a population, with the tradeoff that DNMs can be deleterious and impact the fitness of an individual (Belyeu *et al.*, 2021).

Structural variants (SVs) are altered DNA segments larger than 50 base pairs, that are a change in copy number, chromosomal location or orientation (Escaramís *et al.*, 2015). The detection of *de novo* SVs (dnSVs) is important because they can change gene dosage and structure, and affect gene expression and function by gains or losses of DNA segments (Bhanuprakash *et al.*, 2018; Ran *et al.*, 2018). We aimed to detect dnSV in two commercial pig lines using whole genome sequencing of trios and characterise these dnSV, including predicting their causal mechanisms, functional annotations, and where possible, identifying the parent-of-origin.

Materials & methods

Samples. DNA from 117 individuals (46 trios from two commercial pig lines) was sequenced (Illumina paired-end sequencing, mean coverage = 30X) and aligned to the Sscrofa11.1 assembly (NCBI, 2017). Nine trios were excluded because of poor DNA sequence quality.

SV detection and de novo candidate filtering. We used a Smoove (v.0.2.8) pipeline (v.0.1.0) for calling SVs (Pita Barros, 2021). Three types of SVs were identified for our study: deletions, duplications, and inversions. We filtered SVs using Bcftools (v.1.9), VCFtools (v.4.0.0) and custom R (v.3.6.2) and Python

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

DOI: 10.3920/978-90-8686-940-4_562, © M. Steensma *et al.* 2022

(v.2.7.15) scripts. SV were declared a dnSV when the proband was heterozygous for the SV and parents and unrelated trios did not have the SV. The sum of genotype quality (GQ) scores for a trio had to be greater than 120. We included additional filters based on SV type, checking depth fold-change, minimum and maximum allelic balance for probands and parents respectively, and a maximum of three reads supporting the dnSV were allowed in either parent. Finally, candidate sites were manually inspected using Integrative Genome Viewer (IGV) (Robinson *et al.*, 2011), to remove spurious dnSV cases.

Characterising dnSVs. To identify the parent-of-origin, we looked at informative SNPs located within the dnSVs. For deletions, the informative SNPs we used were homozygous for the alternative allele in the proband, and homozygous for the reference allele in one of the parents and for duplications, informative SNPs were heterozygous with a 2:1 ratio and one parent homozygous (for either allele) and the other heterozygous (Brandler *et al.*, 2016). Genes located within the dnSVs were retrieved from Ensembl (release 104, May 2021) and NCBI annotation (release 106, May 2017) using NCBI Genome Data Viewer (Rangwala *et al.*, 2021). We identified the most likely causal mechanisms for dnSV formation based on split reads spanning the breakpoints of the dnSVs. Based on the degree of sequence homology at junctions we classified them as either non-homology (Belyeu *et al.*, 2021) or microhomology (Hastings *et al.*, 2009; Quinlan *et al.*, 2010; Yan *et al.*, 2007).

Results

Identified dnSVs. A total of 90,031 structural variants (SVs), including 46,478 deletions, 6,541 duplications, 5,977 inversions, and 31,035 breakend class variants (not included in further analyses) were identified. After filtering for dnSVs, we retained 187 candidate dnSVs and after visual inspection, we identified four dnSVs (Table 1). We estimated a mutation rate of 0.108 per generation (1 dnSV per 9 offspring). We identified two DNM clusters. The first involved a 187-bp duplication, a *de novo* single nucleotide variant within the duplication and a small 11-bp deletion at the distal breakpoint of the duplication. The second was a complex dnSV cluster, which consisted of one deletion and three duplications, of which one duplication was inverted.

Characterising dnSVs. The complex dnSV cluster had informative SNPs in each of the four components identifying it as a paternal haplotype. The other dnSVs had no informative SNPs located within the breakpoints. All gene annotations and predicted mutational mechanisms are reported in Table 1.

Table 1. *De novo* structural variants detected in two pig lines using whole genome sequencing of trios.

dnSV type	Position	Mechanism	Gene
Duplication	Chr1:2940197-2940384	Non-homology	[1]Paralogue of *AGP7*[2]
Complex cluster			
- Deletion	Chr1:95207720-95210792	Micro-homology	*SLC14A2*[2]
- Duplication	Chr1:95210182-95213638	Complex	
- Duplication	Chr1:95213834-95214842	Non-homology	
- Duplication	Chr1:95214800-95232460	Micro-homology	
Duplication	Chr7:37837360-37837933	Non-homology	*BICRAL*[3]
Deletion	Chr15:18965177-18965241	Non-homology	*NCKAP5*[2]

[1] *LOC110256592*.
[2] Source: NCBI.
[3] Source: Ensembl.

Discussion

This is the first study to detect and characterise dnSV in pigs. The dnSV rate, 0.108, was similar to that found in humans Belyeu *et al.* (2021) (0.122). Some dnSVs could remain undetected because it is challenging to uniquely map short-reads to repetitive DNA (Weckselblatt and Rudd, 2015). The use of more costly long-read sequencing technologies would span these repetitive regions and aid in dnSV detection there. Regardless, the DNMs identified in this could contribute to the genetic improvement programs of the evaluated commercial lines in this study, if future studies show they affect production and functional traits.

Complex mutation clusters. Two of the four high evidence dnSVs co-occurred with other DNMs and were complex mutational clusters. The observation of mutational clusters is consistent with other studies (Brandler *et al.*, 2016; Hermetz *et al.*, 2014). A possible explanation are mutational hotspots, which means we would expect an increased number of standing variants near the breakpoints of dnSVs too (Brandler *et al.*, 2016). Understanding these complex mutation clusters is necessary to fully interpret the effects on phenotypic characteristics.

Mutational mechanisms. We found that three out of four high evidence dnSVs were mediated by mechanisms that do not require sequence homology at breakpoints. This agreed with the study by Belyeu *et al.* (2021), which showed that 75% of the detected dnSVs in humans lacked sequence homology. We found that the dnSV cluster arose through several mechanisms, including non-homology, micro-homology and an unidentified complex mechanism. In other SV studies classifications are more specific (Hastings, Lupski, *et al.*, 2009), however, identifying the exact causal mechanisms is complex, and mechanisms can act together (Hastings *et al.*, 2009).

Intronic dnSVs and gene expression. We found that the four dnSVs were overlapping within the first introns of protein-coding genes, where regulatory elements are known to be preferentially located (Rigau *et al.*, 2019). The three known genes, presented in Table 1, are: (*BICRAL)* associated with the Coffin-Siris Syndrome 3 in humans which causes congenital malformation including feeding difficulties and poor growth; (*NCKAP5*) associated with Attention Deficit-Hyperactivity Disorder and Drug-Induced Lupus Erythematosus in humans; and (*SLC14A*) involved in the urea transport family (GeneCards). However, the effects of the detected dnSVs on the pigs they were identified in, remains to be discovered.

Ethics statement

All biological material used in this study was collected as part of routine commercial operations.

Funding

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri and Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. This research is supported by the NWO domain TTW.

References

Belyeu J.R., Brand H., Wang H., Zhao X., Pedersen B. S., *et al.* (2021) Am. J. Hum. Genet. 108(4):597–607. https://doi.org/10.1016/j.ajhg.2021.02.012

Bhanuprakash V., Chhotaray S., Pruthviraj D.R., Rawat C., Karthikeyan A., and Panigrahi M. (2018) Vet. World 11(4):535–541. https://doi.org/10.14202/vetworld.2018.535-541

Bishop M.R., Diaz Perez K.K., Sun M., Ho S., Chopra P., *et al.* (2020) Am. J. Hum. Genet.:107(1):124–136. https://doi.org/10.1016/j.ajhg.2020.05.018

Brandler W.M., Antaki D., Gujral M., Noor A., Rosanio G., *et al.* (2016). Am. J. Hum. Genet 98(4):667–679. https://doi.org/10.1016/j.ajhg.2016.02.018

Escaramís G., Docampo E., and Rabionet R. (2015) Brief. Funct. Genomics 14(5):305–314. https://doi.org/10.1093/bfgp/elv014

GeneCards. Available at: https://www.genecards.org/.

Hastings P.J., Ira G., and Lupski J.R. (2009) PLoS Genet 5(1):e1000327. https://doi.org/10.1371/journal.pgen.1000327

Hastings P.J., Lupski J.R., Rosenberg S.M., and Ira, G. (2009) Nat. Rev. Genet. 10(8):551–564. https://doi.org/10.1038/nrg2593

Hermetz K.E., Newman S., Conneely K.N., Martin C.L., Ballif B.C., *et al.* (2014) PLoS Genet. 10(1):e1004139. https://doi.org/10.1371/journal.pgen.1004139

Jin Z.-B., Li Z., Liu Z., Jiang Y., Cai X.-B., and Wu J. (2018) Biol. Rev. 93(2):1014–1031. https://doi.org/10.1111/brv.12383

NCBI. (2017) Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000003025.6/.

Pita Barros, C. (2021). Available at: (https://carolinapb.github.io/population-structural-var-calling-smoove/).

Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. L., Zhang, Y., *et al.* (2010) Genome Res 20(5):623–635. https://doi.org/10.1101/gr.102970.109

R Core Team. (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.r-project.org

Ran X.-Q., Pan H., Huang S.-H., Liu C., Niu X., *et al.* (2018) J. Anim. Physiol. Anim. Nutr. 102(5):1320–1327. https://doi.org/10.1111/jpn.12922

Rangwala S.H., Kuznetsov A., Ananiev V., Asztalos A., Borodin E., *et al.* (2021) Genome Res. 31(1):159–169. https://doi.org/10.1101/gr.266932.120

Rigau M., Juan D., Valencia A., and Rico D. (2019) PLoS Genet. 15(1):e1007902. https://doi.org/10.1371/journal.pgen.1007902

Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., *et al.* (2011) Nat. Biotechnol. 29(1):24–26. https://doi.org/10.1038/nbt.1754

Sasani T.A., Pedersen B.S., Gao Z., Baird L., Przeworski M., *et al.* (2019) eLife 8:e46922. https://doi.org/10.7554/eLife.46922

Weckselblatt B., and Rudd M.K. (2015) Trends Genet. 31(10):587-599. https://doi.org/10.1016/j.tig.2015.05.010