BMC Genomics

RESEARCH Open Access

Identification and characterisation of *de novo* germline structural variants in two commercial pig lines using trio-based whole genome sequencing

Marije J. 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]

Abstract

Background *De novo* mutations arising in the germline are a source of genetic variation and their discovery broadens our understanding of genetic disorders and evolutionary patterns. Although the number of *de novo* single nucleotide variants (dnSNVs) has been studied in a number of species, relatively little is known about the occurrence of *de novo* structural variants (dnSVs). In this study, we investigated 37 deeply sequenced pig trios from two commercial lines to identify dnSVs present in the offspring. The identified dnSVs were characterised by identifying their parent of origin, their functional annotations and characterizing sequence homology at the breakpoints.

Results We identified four swine germline dnSVs, all located in intronic regions of protein-coding genes. Our conservative, first estimate of the swine germline dnSV rate is 0.108 (95% CI 0.038–0.255) per generation (one dnSV per nine offspring), detected using short-read sequencing. Two detected dnSVs are clusters of mutations. Mutation cluster 1 contains a *de novo* duplication, a dnSNV and a *de novo* deletion. Mutation cluster 2 contains a *de novo* deletion and three *de novo* duplications, of which one is inverted. Mutation cluster 2 is 25 kb in size, whereas mutation cluster 1 (197 bp) and the other two individual dnSVs (64 and 573 bp) are smaller. Only mutation cluster 2 could be phased and is located on the paternal haplotype. Mutation cluster 2 originates from both micro-homology as well as non-homology mutation mechanisms, where mutation cluster 1 and the other two dnSVs are caused by mutation mechanisms lacking sequence homology. The 64 bp deletion and mutation cluster 1 were validated through PCR. Lastly, the 64 bp deletion and the 573 bp duplication were validated in sequenced offspring of probands with three generations of sequence data.

Conclusions Our estimate of 0.108 dnSVs per generation in the swine germline is conservative, due to our small sample size and restricted possibilities of dnSV detection from short-read sequencing. The current study highlights the complexity of dnSVs and shows the potential of breeding programs for pigs and livestock species in general, to provide a suitable population structure for identification and characterisation of dnSVs.

*Correspondence:
Marije J. Steensma
marije.steensma@wur.nl

Full list of author information is available at the end of the article

Keywords *De novo* mutation, Structural variation, Germline mutation, *Sus scrofa*, Trios, Whole genome sequencing, Three-generational, Genomics

Background

De novo mutations (DNMs) are spontaneous mutations in the germline and a source of genetic variation that occur during gametogenesis [1]. The new alterations in the genome are genetic variants absent in the somatic cells of the parents and present in the germline of their offspring. Hereafter offspring where DNMs are detected, are referred to as proband [2]. Depending on the method used for DNM detection, mutations arising in the early fertilized egg cell during embryogenesis, resulting in an individual carrying the mutation in some, but not all tissues and organs, are often mistaken as germline DNM [3, 4]. This phenomenon is known as mosaicism, and depending on when mutations have occurred, mosaicism may be restricted to soma, germline or involving both [5]. When mutations occur in the germline, as in DNMs, they can be transmitted to the next generation. The genetic variation introduced by the mutation allows for continued selection and adaptation in a population, with the tradeoff that DNMs can be deleterious and impact the fitness of an individual [6]. Therefore, identifying DNMs can broaden the understanding of genetic disorders and evolutionary patterns in the short and long term [7].

The origins and patterns of *de novo* single nucleotide variants (dnSNVs) have been studied in humans and other primates [8–10]. In humans, the germline dnSNV rate is estimated at approximately 1×10^{-8} per site per generation, giving rise to 44 to 82 dnSNVs per offspring [4, 11–13]. Several mechanisms are known to cause dnSNVs, mostly involving DNA replication [14]. The male germline undergoes continuous germ cell divisions from puberty onwards, whereas the female germline does not [9]. This has likely resulted in the paternal bias in the origin of dnSNVs that has been found in humans and chimpanzees [9, 11, 15]. Moreover, approximately 2 to 3% of all dnSNVs in offspring were found to occur together (<20 kb) as clustered mutations [8, 16, 17]. These clustered dnSNVs are equally abundant on maternal and paternal gametes and have mutation spectra distinct from non-clustered DNMs, suggesting different underlying mechanisms [8, 18, 19]. Furthermore, studies have revealed dnSNV rates in several species, including birds, cattle, fish and wolves [2, 20–22].

De novo structural variants (dnSVs) are predicted to occur a hundred-fold less frequently than dnSNVs [6]. Structural variants (SVs) are altered DNA segments larger than 50 base pairs (bp), that are a change in copy number (deletion, duplication, insertion), chromosomal location (translocation) or orientation (inversion) [23]. SVs can introduce new changes in gene dosage and structure, and affect gene expression and function by gains or losses of DNA segments [24, 25]. Recent studies have shown that dnSVs are associated with rare genetic disorders in humans, including autism and schizophrenia [3, 6, 26]. Additionally, SVs have been suggested to contribute to the phenotypic variation of economically important traits in livestock species [27–29], emphasizing the importance of detecting dnSVs. However, dnSV rates have only been reported in humans [6] and rhesus macaques [7], with a large variance in estimates of dnSV rates in humans [6]. Contrary to dnSNV detection, the identification of dnSVs in populations remains a major challenge. The detection of SVs is restricted by limited sensitivity of microarray and sequencing-based approaches [30], which has resulted in a limited understanding of the dnSV rate.

The population structure of livestock species is helpful to study DNMs. In contrast to humans, most livestock species, including the domestic pig, have large family sizes and relatively short generation intervals. In addition, their phenotypes, genotypes and relationships are routinely collected in breeding programs, making it possible to investigate the impact of DNMs on the phenotypes of the offspring. While mutation rates in livestock species have not been widely studied, such studies could contribute to our understanding of genetic disorders and evolutionary patterns.

The aim of our study was to detect and characterize dnSVs in two commercial pig lines using trio-based whole genome sequencing (WGS) data. We analysed the rate of DNMs for three major classes of SVs: deletions, duplications and inversions. We provide a first, conservative estimate of swine germline dnSV rate in pigs using short-read WGS data. We also provide fine-scale molecular characterisation of these identified dnSVs, including identification of their parent of origin, functional annotations and characterisation of their sequence homology at the breakpoints.

Results

Candidate dnSVs

A total of 90,031 autosomal SVs, including 46,478 deletions, 6,541 duplications, 5,977 inversions and 31,035 breakend class variants were identified. We found significant positive correlations between the mean sequencing depth and number of duplications ($r=0.620$, $P<0.0001$), inversions ($r=0.405$, $P<0.01$) and break end class variants ($r=0.814$, $P<0.0001$) in line 1, and a significant positive correlation between the mean sequencing depth and the number of breakend class variants ($r=0.593$, $P<0.0001$) in line 2. (Additional file 1: Figure S1). Additionally, a few

samples show an excess of candidate inversions compared to the other samples (Additional file 1: Figure S1). Further analyses focused on deletions, duplications and inversions. Structural variants in the breakend class were not considered because this class contains non-canonical types of SVs which are difficult to map and interpret. After filtering for genotype-based Mendelian inconsistencies, a total of 1,521 deletions, 359 duplications, and 4,082 inversions remained. Six probands had an excess call of candidate *de* novo inversions and all had DNA extracted from semen. A majority of these *de novo* inversions overlap with repeats and were similarly called in multiple probands with an excess of *de novo* inversions (semen samples). Identical dnSVs found in multiple probands are likely false positives, because dnSVs are unique and expected to be a one-time event. In subsequent filtering steps we removed spurious sites based on allele count filters, changes in read-depth, number of reads and identical candidate dnSVs found in multiple probands (see materials and methods). After these filters, we retained 163 candidate dnSVs, including 67 deletions, 15 duplications and 81 inversions in 37 trios that were manually inspected using Integrative Genomics Viewer (IGV) [31].

Identified dnSVs

Based on manual inspection in IGV, we identified four high evidence germline dnSVs in 37 pig trios, ranging between 64 bp and 25 kb (Table 1). We estimated a swine germline dnSV rate of 0.108 (95% CI 0.038–0.255) per generation (one dnSV per nine offspring). Of the four detected dnSVs we identified, two are clustered mutations, consisting of multiple mutation events. Mutation cluster 1 is 197 bp in size and contains a 187 bp *de novo* duplication, a dnSNV within the duplication and an 11 bp *de novo* deletion at the distal breakpoint of the duplication (Fig. 1). Mutation cluster 2 consists of one *de novo* deletion and three *de novo* duplications, of which one duplication is inverted (Fig. 2A). Mutation cluster 2 is 25 kb in size, whereas the size of mutation cluster 1 (197 bp) and the other two detected dnSVs are smaller (64 and 573 bp). IGV screenshots of all identified dnSVs can be found in Additional file 1: Figures S2 to S5. Moreover, we detected a 276 bp mosaic deletion in the proband (Table 1). This mosaic deletion contains heterozygous SNPs showing a 2:1 ratio in the proband and a 1:1 ration in both parents, indicating that ~25% is deleted (Additional file 1: Figure S6). This deletion was determined as mosaic, because it deviates from the 50% deletion expected for true germline dnSVs, and is therefore not included in calculating the dnSV rate nor described further.

For the four dnSVs, we analysed their genic overlap. All identified dnSVs partially overlap with intronic regions of protein-coding genes (Ensembl release 107, July 2022 and NCBI release 106, May 2017) (Table 1): Mutation cluster 1 overlaps with phosphodiesterase 10 A (*PDE10A*), mutation cluster 2 overlaps with the solute carrier family member 2 (*SLC14A2*), the 573 bp duplication overlaps with the BRD4 (Bromodomain Containing 4) interacting chromatin remodeling complex associated protein like (*BICRAL*) and the 64 bp deletion overlaps with the NCK (non-catalytic region of tyrosine kinase adaptor) associated protein 5 (*NCKAP5*).

Validated dnSVs

We searched for the detected dnSVs in the public Ensembl SV database (release 107, July 2022) [32] and a pig-specific variation database, PigVar [33], and found that all four detected dnSVs have not been publicly reported before.

Three out of the four detected dnSVs were validated. We designed PCR primers for three of the four detected dnSVs: (i) the 64 bp deletion, (ii) the 573 bp duplication, (iii) and the 187 bp duplication within mutation cluster

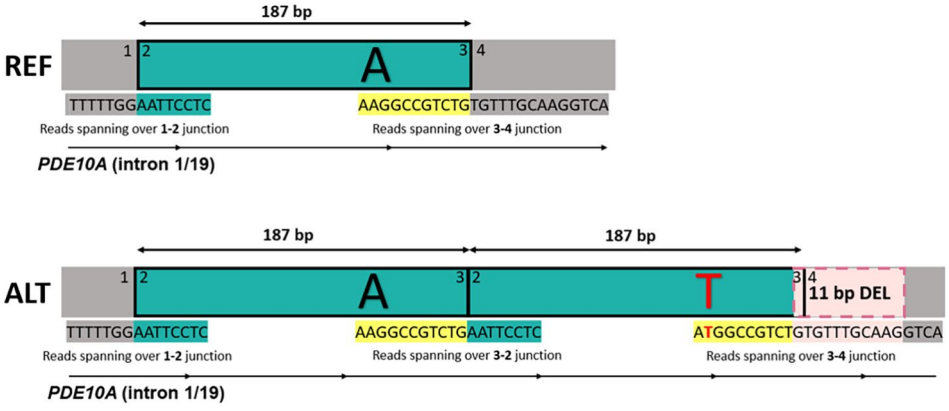

Fig. 1 Reconstruction of mutation cluster 1. REF represents the reference allele. ALT represents the alternate allele involving a 187 bp *de novo* duplication, a *de novo* single nucleotide variant (dnSNV) (A > T) and an 11 bp *de novo* deletion. Reads are shown which are spanning over different junctions (1–4). Intron 1 of the gene *PDE10A* is shown, which overlaps with this region. IGV screenshot of mutation cluster 1 can be found in Additional file 1: Figure S2

Table 1 Identified *de novo* structural variants (dnSVs) from whole genome sequencing (WGS) data of 37 trios

dnSV type	Position	Size (bp)	Parent of origin	Locus	Putative mechanism	PCR √[1]	Offspring √[2]
Mutation cluster 1		197	n.a.	*PDE10A*		Yes	n.a.
-Duplication	Chr1:2940197–2,940,384	187		(intron 1/19)[3]	NON-HOM		
-dnSNV	Chr1:2940375	1			n.a.		
-Deletion	Chr1:2940383–2,940,394	11			n.a.		
Mutation cluster 2		25,196	Paternal	*SLC14A2*		n.a.	n.a.
-Deletion	Chr1:95207720–95,210,792	3072		(intron 3/12)[4]	MICRO-HOM		
-Duplication	Chr1:95210182–95,213,638	3456			Unidentified		
-Duplication	Chr1:95213834–95,214,842	1008			NON-HOM		
-Duplication	Chr1:95214800–95,232,460	17,660			MICRO-HOM		
Duplication	Chr7:37837360–37,837,933	573	n.a.	*BICRAL* (intron 1/11)[3]	NON-HOM	No	Yes
Deletion	Chr15:18965177–18,965,241	64	n.a.	*NCKAP5* (intron 4/17)[4]	NON-HOM	Yes	Yes
Mosaic deletion	Chr11:53586793–53,587,069	276	n.a.	Intergenic	MICRO-HOM	No	n.a.

[1]dnSV validated by PCR.

[2]dnSV validated in one sequenced offspring of the proband.

[3]Source: Ensembl.

[4]Source: NCBI.

n.a.: not available. NON-HOM = non-homology. MICRO-HOM = micro-homology.

1 (Additional file 2: Table S1). Additionally, PCR primers were designed for the 276 bp mosaic deletion (Additional file 2: Table S1). The 187 bp duplication within mutation cluster 1 and the 64 bp deletion were validated to be present in the heterozygous state in the proband and some of its offspring through PCR (Table 1). The 573 bp duplication and the 276 bp mosaic deletion were not validated in proband and its offspring as PCR amplification showed no variation (Table 1, Additional file 2: Table S1).

It would be possible to validate the transmission of germline dnSVs if the proband had sequenced offspring available. The 64 bp deletion and the 573 bp duplication were found in probands where three generations of sequence data were available and the transmission of both dnSVs was validated in the sequenced offspring of the proband (Table 1). We were unable to validate mutation cluster 2 because it was too complex for PCR primer design and the proband did not have any sequenced offspring available (Table 1).

Parent of origin

The presence of informative SNPs located within dnSVs can aid in identifying the parent of origin. Mutation cluster 2, consisting of one deletion and three duplications, has informative SNPs (Additional file 1: Figure S3). The informative SNPs within the deletion of mutation cluster 2 are homozygous for the alternate or reference allele in the proband, and homozygous for the other allele in the father. The informative SNPs within the duplications of mutation cluster 2 are heterozygous with a 2:1 ratio in the proband where the allele with more than expected number of reads (causing the 2:1 ratio) came from the father. Hence, we identified mutation cluster 2 as originating from the paternal gamete (Table 1). The other three identified germline dnSVs had no informative SNPs located within the breakpoints of the dnSVs and therefore the parent of origin could not be determined.

Mutation mechanisms

We were able to categorize the dnSVs by the degree of sequence homology surrounding the breakpoints into two broad categories: non-homology (NON-HOM) (0 to 1 bp) and micro-homology (MICRO-HOM) (2 to 15 bp) (Table 1). Based on these definitions, the 187 bp duplication within mutation cluster 1, the 64 bp deletion, and the 573 bp duplication show no sequence homology at the breakpoints. The deletion and duplications within mutation cluster 2 shows different sequence homology at the breakpoints (Fig. 2B). A 5 bp and a 4 bp micro-homology was found at the breakpoints of the 3,072 bp deletion and 17,660 bp duplication, respectively. No sequence homology was found at the breakpoints of the 1,008 bp duplication and the 3,456 bp inverted duplication. The latter shows a unique feature which was too complex to interpret.

Discussion

The current study focused on identification and characterisation of dnSVs. Here, we report a conservative, first estimate of the swine germline dnSV rate of 0.108 per generation (one dnSV per nine offspring) detected using short-read sequencing data. Our estimate is similar to the rate reported humans (0.122 per generation) [6]. We identified four germline dnSVs, in 37 sire-dam-proband

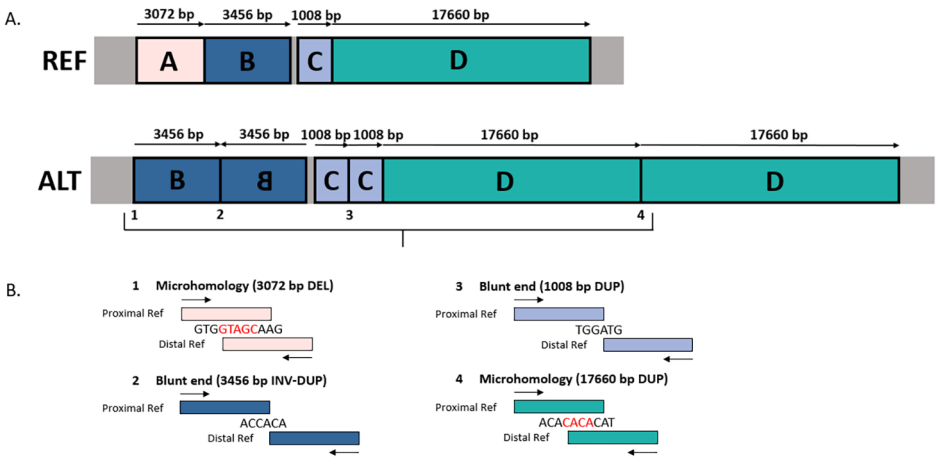

Fig. 2 Reconstruction of mutation cluster 2. A. The reference genome (REF) compared to the alternate mutation cluster 2 (ALT). Mutation cluster 2 represents the structure: Deletion (A) – Inverted duplication (B) – Normal (196 bp) – Duplication (C) – Duplication (D). The 610 bp overlap of deletion A and duplication B and the 42 bp overlap of duplication C and D are not presented in this figure for simplification. The figure is not to scale. **B.** Shows the presence of sequence homology (micro-homology) or no homology (blunt ends) at the junctions of each dnSV. Bases indicated in red represent homology. DEL = deletion, DUP = duplication, INV-DUP = inverted duplication. Integrative Genome Viewer screenshots of mutation cluster 2 can be found in Additional file 1: Figure S3

trios, all of which overlapped with intronic regions of protein-coding genes. Two of the four germline dnSVs are clusters of *de novo* mutations. As in humans [30], clusters of mutations were also observed, where complex mutation clusters consisted of combinations of dnSVs or dnSNVs. We were able to validate two out of three tested dnSVs through PCR, the 64 bp deletion and mutation cluster 1. Additionally, the 64 bp deletion and the 573 bp duplication were found in probands where three generations of sequence data were available and the transmission of both dnSVs were confirmed in the sequenced offspring of the proband. We suggest future studies to use larger cohorts with three generations of long-read sequence data, as it will aid in more accurate detection and validation of germline dnSVs.

Mutation clusters

Two of the four identified dnSVs co-occur with other DNMs and we describe them as mutation clusters. Mutation cluster 1 consists of a *de novo* duplication, a dnSNV and a *de novo* deletion. Mutation cluster 2 consists of one *de novo* deletion and three *de novo* duplications of which one is inverted. The observation of mutational clusters is consistent with a dnSV study in humans, which also found dnSVs consisting of clusters of deletions, duplications and inversions occurring as single events as well as dnSVs co-occurring with dnSNVs [30]. In humans, these mutation clusters could be explained by mutation hotspots, which are regions in the genome where there is an increased number of variants segregating in the population near the breakpoints of dnSVs [30]. A recent study in humans showed that mutation hotspots are enriched with unbalanced dnSVs [34]. Additionally, another study in humans showed that SNVs occurred during the formation of SVs and thereby are enriched in SV hotspots [35]. However, mutation hotspots are species-specific and for pigs yet unknown.

Intronic dnSVs and gene functions

Copy number variations (CNVs; deletions, duplications, insertions) have been shown to capture 17.7% of the total variation in gene expression in humans [36]. The CNVs with large effect size on gene expression were mostly the ones disrupting coding sequence or impacting the regulatory landscape of the region where those CNVs occur [36]. In the current study, we found that all detected dnSVs are located within intronic regions (non-coding sequences) of protein-coding genes. Therefore, we should not necessarily expect these detected dnSVs to contribute to variation in gene expression. Nevertheless, introns encompass half of the non-coding genome and are known to contain important regulatory elements [37]. A study on intronic SVs (deletions, duplications, insertions) in humans found that intronic deletions can result in repressed or enhanced gene expression [37]. Thus, to interpret possible effects of dnSVs it is important to understand the genes with which the dnSVs overlap and their functions.

Previous studies reported functions and diseases associated with the genes overlapping with the dnSVs found in the current study. First, the 573 bp duplication overlaps with the BRD4 (Bromodomain Containing 4) interacting chromatin remodeling complex associated protein like (*BICRAL*), which is a component of the ATP-dependent SWI/SNF (SWItch/Sucrose Non-Fermentable) chromatin remodelling subcomplex GBAF (GLTSCR1 like

containing BRG1/BRM-associated factor). This subcomplex performs important enzymatic activities leading to changes in chromosome structure by altering DNA-histone contacts within a nucleosome [38]. *BICRAL* is associated with Coffin-Siris Syndrome 3 in humans which causes congenital malformation and can cause feeding difficulties and poor growth [39]. *BICRAL* homozygous knockout mice show an embryonic lethal phenotype [40, 41]. Second, mutation cluster 1 overlaps with phosphodiesterase 10 A (*PDE10A*), which plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides. This gene is associated with a hyperkinetic movement disorder in humans [42] and with striatal degeneration in humans [43]. *PDE10A* knockout mice exhibit a resistance to diet-induced obesity and multiple behaviour abnormalities, e.g., decrease in exploratory locomotor activity [41, 44, 45]. Mice homozygous for the *PDE10A* knockout are viable and fertile, although breeding with two homozygote *PDE10A* knockout mice results in reduced litter sizes [44, 45]. Third, the 64 bp deletion overlaps with the NCK (non-catalytic region of tyrosine kinase adaptor) associated protein 5 (*NCKAP5*). This protein-coding gene is uncharacterized but associated with attention deficit-hyperactivity disorder [46] and drug-induced lupus erythematosus in humans [47]. Last, mutation cluster 2 overlaps with the solute carrier family member 2 (*SLC14A2*). This gene is involved in the urea transport family [48] and associated with the following diseases: bone chondrosarcoma [49] and susceptibility to yaws, a primary bacterial infectious disease that infects the skin, bones, and joints [47]. The two genes, *BICRAL* and *PDE10A*, for which the effects in knockout mice have been investigated, only appear if the mice were homozygous for the knock-out allele. In the current study, the pigs carrying the dnSV are all heterozygous for the dnSV, and likely have normal phenotypes since they were selected as breeding candidates for the two commercial pig lines. Nevertheless, homozygous offspring can occur in subsequent generations and might show a more visible effect on the phenotype.

Mutation mechanisms

We found that three out of four detected dnSVs are mediated by mechanisms that do not require sequence homology at the breakpoints. This agrees with a study in humans, which showed that 75% of the detected dnSVs lacked sequence homology at the breakpoints [6]. We found that mutation cluster 2 likely arises through several mechanisms, including non-homology, micro-homology based mechanisms and an unidentified complex mechanism. Other studies have shown that micro-homology at the breakpoints, including micro-homology-mediated break-induced replication, can cause various complex rearrangements [50, 51]. In some other SV studies, classifications of the underlying forces driving dnSVs are more specific [30, 52]. These studies identify classes like non-homologous end joining, micro-homology break-induced replication, mobile element insertion and non-allelic homologous recombination [30, 52]. Because identifying the exact causal mechanisms is complex and mechanisms can act together [53], we only characterized the detected dnSVs based on sequence homology at the breakpoints and grouped them in a non-homology or micro-homology category. We were not able to find macro homology, like non-allelic homologous recombination, at the breakpoints as this is difficult to detect with short-read sequencing data [6]. Thus, to get a better understanding of the underlying mechanisms driving the evolution of dnSVs, a more comprehensive analysis using long-read sequencing approaches is required [54].

Limitations to detect dnSVs with short reads

There are several challenges associated with detecting dnSVs using short-read sequencing, which likely resulted in undetected dnSVs (false negative) and false positive variant calls [55, 56]. In this study, we saw a high rate of false positive variant calls (97.5%) as only four out of the 163 candidate dnSVs (2.5%) were identified as high evidence germline dnSVs. This is mainly due to the excessive number of candidate (*de novo*) inversions in a few samples (Additional file 1: Figure S1), of which the majority overlaps with repeated regions. It is difficult to map short-reads uniquely to repetitive regions [57, 58] and paired-end short-read alignments are not always able to reveal the complete structure of SVs [56]. One approach to overcome these short-read deficiencies, and thereby also lower the rate of false positive and false negative variant calls, is the use of long-read sequencing as the long-reads can span repetitive regions and improve mapping to the reference genome. Long-read sequencing technologies perform better for SV calling than short-read methods [56]. Currently, long-read sequencing is more expensive than short-read sequencing and therefore less used. However, this will ultimately change in the future due to continuous reductions in costs per base using PacBio or Oxford Nanopore sequencing technologies [56]. Therefore, in future dnSV studies, long-read sequencing technologies with adequate read depth will aid in dnSV detection, especially those arising in tandem-repeat sequences which are known to be hypermutable [56].

In this study, we only identified deletions, duplications, and inversions. Other SV types such as translocations and insertions, were not detected due to the software we used (Lumpy) and were assigned to the "breakend" class (breakpoints that cannot be assigned to a certain SV type). In our dataset, ~34% of SVs were assigned to the "breakend" class. The majority of these SVs are repetitive

elements leading to false positive and false negative variant calls [55], but also translocations and insertions were assigned to this class. Thus, dnSV detection is also dependent on the type of SV caller, as several SV callers are designed to detect only certain SV types [55, 56].

Our estimate of the swine germline dnSV rate in the current study (0.108 per generation; one dnSV per nine offspring) is similar to that found in a large human cohort of >4,000 trios (0.122 per generation based on deletions, duplications and inversions, using similar data, methods and criteria) [6]. Nevertheless, our dnSV estimate is preliminary given the small sample size in our study, and likely conservative due to the detection of dnSVs using short-read sequencing data and that we only focused on detection of deletions, duplications, and inversions. Furthermore, the human dnSV study found that nearly 73% of dnSVs originated from paternal gametes [6]. In the current study, we were not able to test for a paternal effect because of the small number of dnSVs detected. From the detected dnSVs, we were only able to determine the parent of origin for one dnSV. The other three dnSVs are smaller in size (64, 197, 573 bp) and do not contain any informative SNPs to determine their parent of origin. Future studies with larger sample sizes and long-read sequencing data could contribute to confirming a parental bias in the origin of dnSVs and more accurate detection of dnSV rate.

Three-generational sequence data

All identified germline dnSVs were found in commercial pig line 1. All trios from line 1 had genomic DNA isolated from ear punch and the trios from line 2 had genomic DNA isolated from ear punch, hair, or semen, where some semen samples showed poor sequencing quality. In line 1, some trios had three-generational sequence data. Two of the four dnSVs were detected in probands with one sequenced third-generation offspring available and transmission of both dnSVs was confirmed. Due to poor sequencing quality of some semen samples and lack of three generational sequence data, it was challenging to confidently detect dnSVs in pig line 2. We were able to validate two out of three PCR tested dnSVs in the proband and some of its offspring. The dnSV which was not validated by PCR, the 573 bp duplication, was, however, validated by confirming transmission in the probands' sequenced offspring. Mutation cluster 2 was not validated as it was too complex for PCR primer design and the proband did not have sequenced offspring available. Additionally, the 276 bp mosaic deletion was not validated by PCR and due to lack of sequenced offspring of the proband, we were not able to derive whether this mosaicism was restricted to the soma or the germline. Thus, three-generational sequence data, including sequenced offspring of the proband will aid in detection and validation of germline dnSVs. We suggest future studies to sequence a larger number of third-generation offspring to increase probabilities of having at least one offspring which inherited the dnSV from the proband [2]. This is more feasible in livestock species compared to humans due to their population structure. In humans, a minimum of four third-generation offspring was shown to be sufficient [6]. However, in livestock, it is feasible to sequence at least eight to ten third-generation offspring, which increases the probability of sequencing at least one offspring that inherited the dnSV from 93.75% (sequencing four third-generation offspring) to 99.6–99.9%. Livestock species, including pigs, provide an opportunity to study dnSVs due to their population structure and routine collecting of data, contributing to a better understanding of genetic disorders and evolutionary patterns.

Conclusions

The current study focused on identification and characterisation of dnSVs in pigs using trio-based whole genome sequence data. We identified four germline dnSVs, all overlapping with intronic regions of protein-coding genes. Two of the four dnSVs are clustered mutations where a dnSV co-occurs with other dnSVs or with a dnSNV. Our estimate of the swine germline dnSV rate is 0.108, which is a conservative estimate, given our small sample size and the challenges of dnSV detection from short reads. The sample size and the dnSVs that are smaller in size in the current study precluded observation of any paternal bias in the origin of dnSVs. Future dnSV studies will benefit from larger cohorts of trios with three-generational long-read sequence data. The current study highlights the complexity of dnSVs and shows the potential of breeding programs for pigs and livestock species in general, to provide a suitable population structure for identification and characterisation of dnSVs.

Methods

Animal samples

The dataset included 117 individuals from 46 sire-dam-proband trios of two commercial pig lines, line 1 (55 samples, constituting 22 trios) and line 2 (62 samples, constituting 24 trios). All trios from line 1 had genomic DNA isolated from ear punch. The trios from line 2 had genomic DNA isolated from ear punch, hair or semen. All trios had two-generation sequence data (WGS of parents and the proband) available. Additionally, some trios of line 1 had three generations of sequence data; four probands were used as sires in other trios, resulting in their offspring having sequence data available. All samples were whole genome sequenced (mean sequencing depth = 32.4X) using Illumina paired-end sequencing, with 150 bp read length and 300 bp fragment length. The paired-end reads were realigned to the pig reference

genome (Sscrofa11.1, GenBank assembly accession number GCF_000003025.6) with the Burrows-Wheeler Aligner (BWA-mem v.0.7.17) [59] to generate a BAM file for each individual. All BAM files were sorted and indexed with SAMtools (v.1.9) [60]. Sequence alignment and variant calling was performed on the High Performance Computing (HPC) cluster at Wageningen University and Research. Nine trios (21 samples) were excluded because of poor data quality including large numbers of discordantly mapped reads.

SV detection

We used a pipeline (v.0.1.0) [61] to perform structural variant (SV) calling in a population using Smoove (v.0.2.8) [62]. This Smoove pipeline used 'Lumpy' (v.0.2.14a) to call SVs in each sample relative to the reference genome. Lumpy uses signals from split reads and discordant paired end reads to predict breakpoints of deletions, duplications, and inversions [63]. Four different types of SVs are identified with Lumpy: deletion, duplication, inversion and breakend variants that cannot be assigned to one of these three classes (therefore ignored in current study) [30]. SVtools (v.0.4.0) was used to combine all SV calls into one single variant call format (VCF) file [64]. Subsequently, all 37 trios were genotyped for population-wide non-redundant SV sites with SVTyper (v.0.7.0), which uses a Bayesian framework that uses allele counts at each junction to determine the likelihood that a genotype is heterozygous or homozygous [65]. The VCF file was re-genotyped with SVTyper to get information for all SVs in all samples for filtering. The population SV pipeline generated a VCF file in which all detected SVs were denoted, and each sample was assigned a genotype for each SV.

De novo candidate filtering

We filtered SVs using BCFtools (v.1.9) [66], VCFtools (v.4.0.0) [67], and custom R (v.3.6.2) [68] and Python (v.2.7.15) [69] scripts. SVs 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. In addition, allele frequencies of dnSVs higher than 0.1 across the whole dataset were excluded, because dnSVs are unique and expected to be a one-time event. We chose this threshold of 0.1, to account for potential dnSV transmissions from proband to sequenced third-generation offspring present in the dataset.To filter out spurious false dnSVs, we included additional filters per SV type. Deletions passed if the median depth inside the dnSV compared to the median depth 1,000 bases left and right of the dnSV (duphold flank fold-change [DHFFC]) [70] was <0.8; average DHFFC of sire and dam was >0.8; minimum allelic balance, fraction of reads supporting the alternate allele out of the reads supporting reference and alternate allele, for proband was >0.05; there were a minimum of three reads supporting the dnSV in the proband; maximum allelic balance for parents was <0.1; a maximum of three reads supporting the dnSV was present in either parent; and the deletion was not called a dnSV in >1 sample.

Duplications passed if the median depth inside the dnSV compared to the median depth of bins with matching GC content (duphold bin fold-change [DHBFC]) [70] was >1.1; average DHBFC of sire and dam was <1.2; minimum allelic balance for proband was >0.1; there was a minimum of three reads supporting the dnSV in the proband; maximum allelic balance for parents was <0.1; a maximum of three reads supporting the dnSV in either parent; and the duplication was not called a dnSV in >1 sample.

Inversions passed if the minimum allelic balance for the proband was >0.2; there were a minimum of five reads supporting the dnSV in the proband; maximum allelic balance for parents was <0.1; there was a maximum of three reads supporting the dnSV in either parent; and the inversion was not called a dnSV in >1 sample.

Finally, candidate sites were manually inspected using Integrative Genomics Viewer (IGV) (release 2.11.9, December 2021) [31] to remove spurious dnSV cases. Manual inspection included some stringent criteria and dnSV cases were considered spurious when (1) both breakpoints of a dnSV were overlapping with repeats, (2) only one side of the dnSV breakpoint was supported with split or paired-end reads, (3) less than three split or paired-end reads supported the dnSVs at each side of the breakpoint, (4) the split or paired-end reads supporting a dnSV breakpoint were not at all overlapping with each other, or (5) there was no clear visual increase (for duplications) or decrease (for deletions) in coverage at both sides of the dnSV breakpoints.

The dnSV rate was estimated by dividing the number of identified true dnSVs with the number of trios analysed. Confidence intervals of the dnSV rate estimate were calculated using the Wilson method [71].

Validation of dnSVs

The detected dnSVs were compared with the public SV database of Ensembl (release 107, July 2022) [32] and with a pig-specific variation database, PigVar [33] (accessed at 12 November 2022), to validate that these identified dnSVs were novel variants not reported before in pigs.

PCR primers were designed to span each of the breakpoints identified in the sequence data. Mutation cluster 2 was too complex for PCR primer design. For the remaining three dnSVs and the mosaic deletion, PCR assays were designed (Additional file 2: Table S1). The primers were tested on DNA from the proband, its offspring, and

its parents. The PCR was performed on 60ng of DNA (6ul), 0.4 µm primer (0.06ul), 2.5ul FirePol 5x Master Mix and 3.5ul MQ. The PCR cycling conditions were 95 °C for 5 min; 35 cycles of 30 s at 95 °C, 45 s at 55 °C annealing temperature, 90 s at 72 °C; followed by a final elongation step of 72 °C for 10 min. The PCR products were loaded on 3% agarose gel.

Furthermore, two dnSVs, the 64 bp deletion and the 573 bp duplication were detected in probands with three generations of sequence data. Hence, the breakpoints of these dnSVs were visualised in sequenced offspring of the corresponding proband using IGV in order to confirm transmission of these dnSVs.

Determining the parent of origin

Presence of informative SNPs located within dnSVs aided in identifying parent of origin. For deletions, the informative SNPs we used were homozygous for the alternate or reference allele in the proband, and homozygous for the other allele in one of the parents. The parent of origin could then be determined to be on the haplotype of the parent who is homozygous reference for the informative SNP alleles [30].

For duplications, informative SNPs were heterozygous with a 2:1 ratio and one parent homozygous (for either allele) and the other heterozygous. The allele with a higher than expected number of reads (causing the 2:1 ratio) indicated the parental haplotype the duplication originated from [30].

Functional annotation

Genes located within the dnSVs were retrieved from Ensembl (release 107, July 2022) and NCBI annotation (release 106, May 2017) using NCBI Genome Data Viewer [72]. GeneCards [73] was used to gain insight in the functional enrichment of genes. The Mouse Genome Database [41] was used to look for phenotypic effects in knockout mice.

Prediction of causal mechanisms

We identified the most likely causal mechanisms for dnSV formation based on split reads spanning the breakpoints of the dnSVs. The dnSVs were grouped into two broad categories based on the degree of sequence homology at the junctions, using methodology from a similar analysis of dnSVs in human families [6]. dnSVs which showed no breakpoint homology (0 to 1 bp) were grouped as the non-homology based category (NON-HOM), of which non-homologous-end-joining is most often the cause of this type of mechanism [6]. Additionally, dnSVs which showed 2 to 15 bp sequence homology flanking the breakpoints were grouped as the micro-homology based category (MICRO-HOM) [74]. Two main mechanisms of this class are micro-homology-mediated break-induced replication and microhomology-mediated end joining [52, 74, 75].

List of abbreviations

DNM	*de novo* mutation
dnCNV	*de novo* copy number variant
dnSNV	*de novo* single nucleotide variant
dnSV	*de novo* structural variant
MICRO-HOM	micro-homology
NON-HOM	non-homology
SNP	single nucleotide polymorphism
VCF	variant call format
WGS	whole genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09296-3.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors acknowledge Hendrix Genetics B.V. (Boxmeer, the Netherlands) and Topigs Norsvin (Beuningen, the Netherlands) for providing data collection for whole genome sequence data. The use of the HPC Anunna cluster has been made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR).

Author contributions

MS drafted the manuscript under supervision of YL, AB, RC, MD, MG and CR. MS performed the analyses under supervision of YL and CR. CB developed the structural variant calling pipeline used in this study, and CB and MD supervised MS during the use of this pipeline. RC performed PCR validation of the dnSVs. HM, BH, MB, AH and CR contributed to data collection and conception of the study. AB, MG, CR, YL, RC and MS participated in discussions. All authors read and approved the final manuscript.

Funding

This study was conducted within the STW-Breed4Food Partnership, project number 14297: The contribution of *de novo* mutations to long-term selection response in genomic breeding programs (Mutabreed). This study was financially supported by NWO-TTW and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The use of the HPC cluster has been made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR).

Data availability

The data that support the findings of this study are available from Hendrix Genetics B.V. and Topigs Norsvin but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Hendrix Genetics B.V. and Topigs Norsvin.

Declarations

Ethics approval and consent to participate

All biological material used in this study was collected as part of routine data collection from Hendrix Genetics and Topigs Norsvin breeding programs, and not specifically for the purpose of this project. Therefore, approval of an ethics committee was not mandatory. Sample collection was conducted strictly in line with Dutch law on the protection of animals (Gezondheids en welzijnswet voor dieren).

Consent for publication

Not applicable.

Competing interests
The authors declare that this study received funding from Cobb Europe, CRV B.V., Hendrix Genetics B.V. and Topigs Norsvin. All funders were involved in the study design, Hendrix Genetics B.V. and Topigs Norsvin performed data collection and were involved in preparation of the manuscript. The Breed4Food partners Cobb Europe, CRV B.V., Hendrix Genetics B.V. and Topigs Norsvin, declare to have no competing interests for this study. MG is a member of the editorial board (Associate Editor) of BMC Genomics Journal. All authors declare that the results are presented in full and as such present no conflict of interest.

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

Received: 30 January 2023 / Accepted: 4 April 2023
Published online: 18 April 2023

References

1. Bishop MR, Perez KKD, Sun M, Ho S, Chopra P, Mukhopadhyay N, et al. Genome-wide enrichment of de novo coding mutations in orofacial cleft trios. Am J Hum Genet. 2020;107:124–36.
2. Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M et al. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle.BioRxiv. 2017:079863.
3. Jin Z, Li Z, Liu Z, Jiang Y, Cai X, Wu J. Identification of de novo germline mutations and causal genes for sporadic diseases using trio-based whole - exome/genome sequencing. Biol Rev. 2018;93:1014–31.
4. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife. 2019;8:e46922.
5. Cao Y, Tokita MJ, Chen ES, Ghosh R, Chen T, Feng Y, et al. A clinical survey of mosaic single nucleotide variants in disease-causing genes detected by exome sequencing. Genome Med. 2019;11:1–11.
6. Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. Am J Hum Genet. 2021;108:597–607.
7. Thomas GWC, Wang RJ, Nguyen J, Alan Harris R, Raveendran M, Rogers J, et al. Origins and long-term patterns of copy-number variation in rhesus macaques. Mol Biol Evol. 2021;38:1460–71.
8. Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, et al. Parent-of-origin-specific signatures of de novo mutations. Nat Genet. 2016;48:935–9.
9. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488:471–5.
10. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, et al. An extended genotyping framework for Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun. 2016;7:1–11.
11. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends Genet. 2013;29:575–84.
12. Lindsay SJ, Rahbari R, Kaplanis J, Keane T, Hurles ME. Similarities and differences in patterns of germline mutation between mice and humans. Nat Commun. 2019;10:1–12.
13. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13:565–75.
14. Crow JF. The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet. 2000;1:40–7.
15. Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, McVean G. Strong male bias drives germline mutation in chimpanzees. Science (1979). 2014;344:1272–5.
16. Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, et al. Multi-nucleotide de novo mutations in humans. PLoS Genet. 2016;12:e1006315.
17. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151:1431–42.
18. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015;47:822–6.
19. Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Neerincx PB, Bodian DL, et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. Nat Genet. 2018;50:487–92.
20. Feng C, Pettersson M, Lamichhaney S, Rubin C-J, Rafati N, Casini M, et al. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. Elife. 2017;6:e23907.
21. Koch EM, Schweizer RM, Schweizer TM, Stahler DR, Smith DW, Wayne RK, et al. De novo mutation rate estimation in wolves of known pedigree. Mol Biol Evol. 2019;36:2536–47.
22. Smeds L, Qvarnström A, Ellegren H. Direct estimate of the rate of germline mutation in a bird. Genome Res. 2016;26:1211–8.
23. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14:305–14.
24. Bhanuprakash V, Chhotaray S, Pruthviraj DR, Rawat C, Karthikeyan A, Panigrahi M. Copy number variation in livestock: a mini review. Vet World. 2018;11:535.
25. Ran X, Pan H, Huang S, Liu C, Niu X, Li S, et al. Copy number variations of MTHFSD gene across pig breeds and its association with litter size traits in chinese indigenous Xiang pig. J Anim Physiol Anim Nutr (Berl). 2018;102:1320–7.
26. Rees E, Kirov G, O'Donovan MC, Owen MJ. De novo mutation in schizophrenia. Schizophr Bull. 2012;38:377–81.
27. Liu GE, Bickhart DM. Copy number variation in the cattle genome. Funct Integr Genomics. 2012;12:609–24.
28. Wang Y, Gu X, Feng C, Song C, Hu X, Li N. A genome-wide survey of copy number variation regions in various chicken breeds by array comparative genomic hybridization method. Anim Genet. 2012;43:282–9.
29. Bickhart DM, Liu GE. The challenges and importance of structural variation detection in livestock. Front Genet. 2014;5:37.
30. Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, et al. Frequency and complexity of de novo structural mutation in autism. Am J Hum Genet. 2016;98:667–79.
31. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
32. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.
33. Zhou Z-Y, Li A, Otecko NO, Liu Y-H, Irwin DM, Wang L et al. PigVar: a database of pig variations and positive selection signatures. Database. 2017;2017.
34. Lin Y-L, Gokcumen O. Fine-scale characterization of genomic structural variation in the human genome reveals adaptive and biomedically relevant hotspots. Genome Biol Evol. 2019;11:1136–51.
35. Abyzov A, Li S, Kim DR, Mohiyuddin M, Stütz AM, Parrish NF, et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nat Commun. 2015;6:1–12.
36. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science (1979). 2007;315:848–53.
37. Rigau M, Juan D, Valencia A, Rico D. Intronic CNVs and gene expression variation in human populations. PLoS Genet. 2019;15:e1007902.
38. Alpsoy A, Dykhuizen EC. Glioma tumor suppressor candidate region gene 1 (GLTSCR1) and its paralog GLTSCR1-like form SWI/SNF chromatin remodeling subcomplexes. J Biol Chem. 2018;293:3892–903.
39. Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat Genet. 2012;44:376–8.
40. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016;537:508–14.
41. Blake JA, Baldarelli R, Kadin JA, Richardson JE, Smith CL, Bult CJ. Mouse Genome Database (MGD): knowledgebase for mouse–human comparative biology. Nucleic Acids Res. 2021;49:D981–7.

42. Diggle CP, Rizzo SJS, Popiolek M, Hinttala R, Schülke J-P, Kurian MA, et al. Biallelic mutations in PDE10A lead to loss of striatal PDE10A and a hyperkinetic movement disorder with onset in infancy. Am J Hum Genet. 2016;98:735–43.
43. Mencacci NE, Kamsteeg E-J, Nakashima K, R'Bibo L, Lynch DS, Balint B, et al. De novo mutations in PDE10A cause childhood-onset chorea with bilateral striatal lesions. Am J Hum Genet. 2016;98:763–71.
44. Siuciak JA, McCarthy SA, Chapin DS, Fujiwara RA, James LC, Williams RD, et al. Genetic deletion of the striatum-enriched phosphodiesterase PDE10A: evidence for altered striatal function. Neuropharmacology. 2006;51:374–85.
45. Siuciak JA, McCarthy SA, Chapin DS, Martin AN, Harms JF, Schmidt CJ. Behavioral characterization of mice deficient in the phosphodiesterase-10A (PDE10A) enzyme on a C57/Bl6N congenic background. Neuropharmacology. 2008;54:417–27.
46. Bogari NM, Al-Allaf FA, Aljohani A, Taher MM, Qutub NA, Alhelfawi S, et al. The co-existence of ADHD with autism in saudi children: an analysis using next-generation DNA sequencing. Front Genet. 2020;11:548559.
47. Genecards. Genecards - the human gene database. Weizman Instute of Science. 1997. https://www.genecards.org/. Accessed 21 Jan 2022.
48. Smith CP, Fenton RA. Genomic organization of the mammalian SLC14a2 urea transporter genes. J Membr Biol. 2006;212:109–17.
49. Aroankins TS, Murali SK, Fenton RA, Wu Q. The Hydrogen-Coupled Oligopeptide Membrane Cotransporter Pept2 is SUMOylated in Kidney Distal Convoluted Tubule Cells.Front Mol Biosci. 2021;8.
50. Bahrambeigi V, Song X, Sperle K, Beck CR, Hijazi H, Grochowski CM, et al. Distinct patterns of complex rearrangements and a mutational signature of microhomeology are frequently observed in PLP1 copy number gain structural variants. Genome Med. 2019;11:1–17.
51. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet. 2009;41:849–53.
52. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009;5:e1000327.
53. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10:551–64.
54. Imaizumi T, Yamamoto-Shimojima K, Yanagishita T, Ondo Y, Yamamoto T. Analyses of breakpoint junctions of complex genomic rearrangements comprising multiple consecutive microdeletions by nanopore sequencing. J Hum Genet. 2020;65:735–41.
55. Cameron DL, di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019;10:3240.
56. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:1–14.
57. Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. Trends Genet. 2015;31:587–99.
58. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. Nature. 2020;583:83–9.
59. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
60. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
61. Barros CP. Available at: (https://carolina.github.io/population-structural-var-calling-smoove/). 2021.
62. Pedersen BS, Layer R, Quinlan AR. Smoove: structural-variant calling and genotyping with existing tools. 2020.
63. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:1–19.
64. Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, et al. Svtools: population-scale analysis of structural variation. Bioinformatics. 2019;35:4782–7.
65. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12:966–8.
66. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33:2037–9.
67. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
68. R Core Team. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2022.
69. vanRossum G. Python reference manual.Department of Computer Science [CS]. 1995; R 9525.
70. Pedersen BS, Quinlan AR. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. Gigascience. 2019;8:giz040.
71. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat. 1998;52:119–26.
72. Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, et al. Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV). Genome Res. 2021;31:159–69.
73. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics. 2016;54:1–30.
74. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 2010;20:623–35.
75. Yan CT, Boboila C, Souza EK, Franco S, Hickernell TR, Murphy M, et al. IgH class switching and translocations use a robust non-classical end-joining pathway. Nature. 2007;449:478–82.

Publisher's note