Structural variations in pig genomes

Yogesh Paudel

Thesis committee

Promotor
Prof. Dr M.A.M. Groenen
Personal chair at the Animal Breeding and Genomics Centre
Wageningen University

Co-promotor
Dr O. Madsen
Research Associate at the Animal Breeding and Genomics Centre
Wageningen University

Dr H-J. Megens
Research Associate at the Animal Breeding and Genomics Centre
Wageningen University

Other members
Prof. Dr D. de Ridder, Wageningen University
Dr R. van Ham, Keygene N.V., Wageningen, the Netherlands
Dr A. Schönhuth, Center for Mathematics and Computer Science, Amsterdam, the Netherlands
Dr C. Alkan, Bilkent University, Ankara, Turkey

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Structural variations in pig genomes

Yogesh Paudel

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday January 20, 2015
at 4. p.m. in the Aula.

Paudel, Y.
Structural variations in pig genomes,
204 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2015)
With references, with summaries in English and Dutch

ISBN 978-94-6257-214-0

Abstract

Paudel, Y. (2015). Structural variations in pig genomes. PhD thesis, Wageningen University, the Netherlands

Structural variations are chromosomal rearrangements such as insertions-deletions (INDELs), duplications, inversions, translocations, and copy number variations (CNVs). It has been shown that structural variations are as important as single nucleotide polymorphisms (SNPs) in regards to phenotypic variations. The general aim of this thesis was to use next generation sequencing data to improve our understanding of the evolution of structural variations such as CNVs, and INDELs in pigs. We found that: 1) the frequency of copy number variable regions did not change during pig domestications but rather reflected the demographic history of pigs. 2) CNV of olfactory receptor genes seems to play a role in the on-going speciation of the genus *Sus*. 3) Variation in copy number of olfactory receptor genes in pigs (*Sus scrofa*) seems to be shaped by a combination of selection and genetic drift, where the clustering of ORs in the genome is the major source of variation in copy number. 4) Analysis on short INDELs in the pig genome shows that the level of purifying selection of INDELs positively correlates with the functional importance of a genomic region, i.e. strongest purifying selection was observed in gene coding regions. This thesis provides a highly valuable resource for copy number variable regions, INDELs, and SNPs, for future pig genetics and breeding research. Furthermore, this thesis discusses the limitations and improvements of the available tools to conduct structural variation analysis and insights into the future trends in the detection of structural variations.

Contents

1

General introduction

1.1 Introduction

Structural variations (SVs) are rearrangements in a genome such as insertions, deletions, inversions, translocations, and copy number variations. SVs can encompass millions of bases of DNA, containing genes and regulatory regions (Sebat et al. 2004; Iafrate et al. 2004; Tuzun et al. 2005; Redon et al. 2006). Establishing a link between these SVs and phenotypic variations is a challenging job for present-day genome research. While studies have found drastic effects of single nucleotide polymorphisms (SNPs) on phenotypes (Hoekstra et al. 2006; Kijas et al. 2012), SNPs alone will not explain all the existing phenotypic diversity at inter and intra-specific levels. Recent studies have generated high-resolution SV databases and have shown that genomic variations other than SNPs play a prominent role in diseases, complex traits, and evolution (Redon et al. 2006; Perry et al. 2007; Conrad et al. 2006; Sudmant et al. 2010; Mills et al. 2011; Dennis et al. 2012; Durkin et al. 2012; Montgomery et al. 2013).

Based on differences in copy number of affected segments of DNA between/within populations, SVs are divided into two main classes: unbalanced and balanced. Copy number variations (CNVs) and segmental duplications are examples of unbalanced SVs caused by insertion, deletion, and duplication events in a genome, where the number of copies of a segment of DNA varies between/within populations. On the other hand, inversions and translocations are examples of balanced SVs where the number of copies of SV affected segments remain the same between/within populations. Different SV formation mechanisms play a role in the generation of different types of SVs (Mills et al. 2011; Gokcumen et al. 2013; Pang et al. 2013). A comprehensive map of SVs in a genome is essential to understand their role in relation to different phenotypes. Because of the variation in size and occurrence in the genome, and the unclear mechanism in the formation of SVs, the identification of SVs has been a challenge.

1.2 Mechanisms generating structural variation

Systematic and comprehensive estimation of SVs has been problematic and has remained difficult, as the mechanisms that result in SVs are still not well understood. Recently, three major DNA repair mechanisms have been proposed that could be responsible for most of the rearrangement events in mammalian genomes. Two of the mechanisms are based on recombination: non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ). The third, fork stalling and template switching (FoSTeS), is based on replication (Figure 1.1) (Critchlow and Jackson 1998; van Gent et al. 2001; Inoue and Lupski 2002; Yu and Lieber 2003; Lupski 2004; Lee et al. 2007; Gu and Lieber 2008). These rearrangement mechanisms facilitated by DNA repair events probably account for the majority of the SVs (Kidd et al. 2008; Mills et al. 2011).

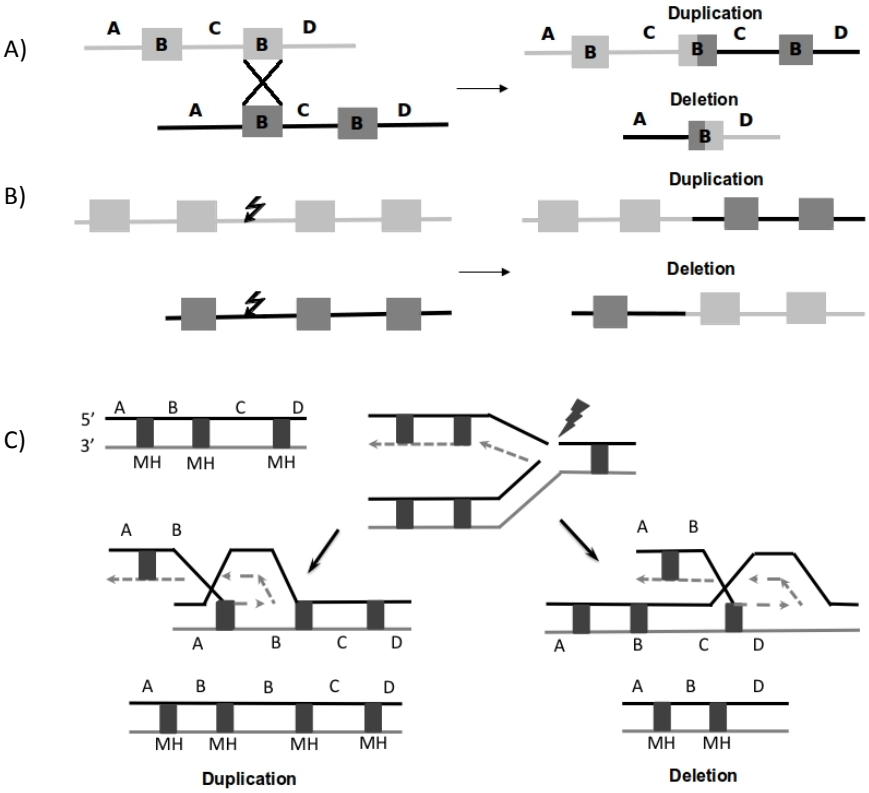


Figure 1.1 A) Non-allelic homologous recombination (NAHR) showing an unequal crossing over between flanking non-allelic homologous sequence blocks (noted as B), which results in reciprocal deletion and duplication. B) Non-homologous end-joining (NHEJ) double-strand break repair mechanism where the ends of DNA double-strand break is joined without any homologous fragments. C) Fork stalling and template switching (FoSTeS) where, the DNA replication fork breaks off, the lagging strand (5') disengages from the original template and switches to another replication fork via micro-homology (MH) and restarts DNA synthesis on the new fork and can result in deletion and or duplication of that sister chromosome.

NAHR occurs due to the alignment and subsequent crossover between two highly similar non-allelic DNA sequences (Figure 1.1A) (Inoue and Lupski 2002; Lupski 2004). It has been suggested that repeats in the immediate vicinity and in the same orientation on the same chromosome mediate duplication and/or deletion (Edelmann et al. 1999; Shaw et al. 2002). Different studies have found an enrichment of highly repeated elements around the flanking regions of CNVs and/or segmental duplications, supporting the importance of NAHR as a prevalent mechanism for the formation of CNVs (Redon et al. 2006; Kidd et al. 2008; Sudmant et al. 2010; Bickhart et al. 2012; Paudel et al. 2013). NAHR between genomic intervals flanked by inverted repeats are suggested to create inversions (Lupski 1998; Stankiewicz and Lupski 2002; Carvalho et al. 2011). Similarly, NAHRs between sequences on different chromosomes help to create chromosomal translocations (Lupski 1998; Stankiewicz and Lupski 2002).

NHEJ is another DNA repair mechanism that aims to repair DNA double strand breaks caused by ionizing radiation or reactive oxygen (van Gent et al. 2001; Yu and Lieber 2003; Agarwal et al. 2006; Gu and Lieber 2008). Like NAHR, NHEJ does not need homologous DNA segments near the breakpoints (Figure 1.1B). NHEJ is considered an imperfect DNA repair mechanism as in most cases it causes either a deletion or an insertion of several nucleotides (Gu and Lieber 2008; Lieber 2010). Breakpoints of NHEJ-mediated genomic rearrangements often occur in the vicinity of repetitive elements such as Alu, LINE, LTR, MIR, and MER2 DNA elements (Nobile et al. 2002; Toffolatti et al. 2002).

The third DNA repair mechanism, fork stalling and template switching (FoSTeS), is a repair mechanism that is induced by single strand breaks during the DNA replication process (Lee et al. 2007). This repair mechanism does not require extensive sequence homology (Figure 1.1C). In this mechanism, "the DNA replication fork can stall, the lagging strand disengages from the original template and switches to another replication fork and restarts DNA synthesis on the new fork by priming it via the micro-homology between the switched template site and the original fork" (Lee et al. 2007). Among the three DNA repair mechanisms, NAHR accounts for most of the rearrangements in a genome (Gu et al. 2008).

1.3 Implication of structural variation in disease

Structural variations comprise a considerable proportion of variation among individuals within a species/population and have been found to influence disease phenotypes by altering dosage sensitive genes, disrupting functional genes, and other molecular mechanisms (Bassett 1998, 2003; Antshel 2007; Salmon Hillbertz et al. 2007; Wright et al. 2009; Mefford et al. 2010; Brouwers et al. 2012). In the majority of cases, SVs have been found to be benign, resulting in only minor or no phenotypic variation (Giuffra et al. 2002; Dumas et al. 2007; Perry et al. 2007; Dennis et al. 2012). However, recent studies on human diseases have provided insight into the functional impact of SVs by associating SVs with complex traits such as autism (Antshel 2007; Eliez 2007), schizophrenia (Bassett 1998, 2003; Arinami 2006), Parkinson (Pankratz et al. 2009), Alzheimer (Brouwers et al. 2012), and epilepsy (Mefford et al. 2010). SVs, especially CNVs, have not only been found to be associated with diseases but also with susceptibility or resistance to different complex traits/syndromes such as AIDS (Gonzalez et al. 2005), Crohn disease (Ogura et al. 2001; Parkes et al. 2007), glomerulonephritis (Fanciulli et al. 2007), and psoriasis (Huffmeier et al. 2010).

1.4 Impact of structural variation on phenotypic traits in domestic animals

Generations of selective breeding of species such as cattle, horse, goat, sheep, dog, and pig for certain traits of interest has resulted in many different varieties or breeds. This process of artificial selection of certain traits of an animal that ultimately benefits the interests of humans is called domestication. Interest in SV detection has recently been extended to domesticated animals to understand the impact of SVs on genomes, which causes variation in phenotypes in these animals (Fadista et al. 2008; Nicholas et al. 2009; Chen et al. 2009b; Bae et al. 2010; Fadista et al. 2010; Fontanesi et al. 2010; Liu et al. 2010; Ramayo-Caldas et al. 2010; Alvarez and Akey 2011; Bickhart et al. 2012; Kijas et al. 2012; Esteve-Codina et al. 2013; Paudel et al. 2013). Some of these studies suggest a role for SVs in several important phenotypic traits in animals that were preferentially selected during the domestication and subsequent breeding process. For example, white coat color in some widely used pig breeds, is caused by a duplication involving the *KIT* gene (Wiseman 1986; Giuffra et al. 2002). The high copy number of amylase genes in domesticated dogs, compared to its wild counterpart, lead to adaptation to food that is rich in starch (Axelsson et al. 2013). The dorsal hair ridge phenotype in dogs (due to the duplication of the *FGF3*, *FGF4*, *FGF19* and *ORAOV1* genes) is another example of the effect of genomic SVs, which were selected in some domestic dog breeds (Salmon Hillbertz et al. 2007). The peacomb phenotype of chicken (reduction of the size of comb and wattles), an adaptive trait in cold climates as it reduces heat loss and makes chicken less susceptible to frost, has been linked to a duplication near the *SOX5* gene (Wright et al. 2009). Another example in chicken is the partial duplication of the *PRLR* and *SPEF2* genes at the late feathering locus which causes a delay in the emergence of flight feathers at hatch (Elferink et al. 2008). These examples demonstrate that the genomic SVs can have phenotypic consequences associated with traits beneficial for humans and positively selected during domestication.

1.5 Impact of structural variation on genome evolution and speciation

Structural variations such as CNVs can play a role in creating new functions for genes, altering gene dosage, reshaping gene structures, and/or modifying the regulatory elements that control gene expression (Long 2001; Otto and Yong 2002; Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010; Dennis et al. 2012). Therefore, understanding the evolution of genomic SVs is vital for understanding how SVs contribute to the evolution of an organism (Long 2001; Otto and Yong 2002; Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010). Dumas et al. observed a higher rate of copy number gain regions encompassing genes compared to copy number losses in primates and proposed that positive selection is involved to explain this observation (Dumas et al. 2007). The authors further suggested that studies on human lineage specific CNVs, may reveal the evolutionary process driving the emergence of human-specific traits such as cognition (Dumas et al. 2007). Recently, Dennis et al., (2012) have identified a region containing the *SRGAP2* gene in the human genome, which was partially duplicated around three million years ago (mya) thereby creating a novel gene function associated with cognitive abilities in humans. Another region in the human genome shows a SV that overlaps with *AQP7*, a gene whose protein is involved in the transport of water and glycerol. SV in human at this region suggests positive selection for thermoregulation by increasing of sweating in human, an important human specific trait (Dumas et al. 2007). Similarly, the salivary amylase gene, *AMY1*, which is positively correlated with the levels of salivary amylase protein and the amount of starch in the human diet, has also been found positively selected in different human populations (Perry et al. 2007). In addition, in other organisms such as flies (*Drosophila melanogaster*), a positive selection of CNV gain regions has been observed. This CNV region encompasses a gene involved in toxin-response (*Cyp6g1*), contributing to a resistance to DTT (Emerson et al. 2008).

These examples of species specific gene duplication and positive selection of specific regions further support the hypothesis that SVs encompassing functional genes can be evolutionarily favored because of their adaptive value. Even though the importance of SVs in speciation, particularly inversions, has been demonstrated through detailed studies in flies (reviewed by (Noor et al. 2001)), the overall role of other types of SVs in the process of speciation is still not clear. Most importantly, the role of SVs in the process of speciation is another unexplored topic hindered by the lack of data from evolutionarily closely related species in which speciation is still ongoing.

1.6 Structural variation detection

1.6.1 Cytogenetic methods

Studies to detect SVs at the chromosomal level already started in the early 20th century using cytogenetic approaches (Sturtevant 1920). Fluorescent In Situ Hybridization (FISH) is an example of a cytogenetic approach developed in the early 1980s, which is still widely used to detect SVs (Langer-Safer et al. 1982). FISH is an experimental protocol that has been used to detect not only the presence or absence of specific DNA sequences on chromosomes but also to estimate the quantity and location of those regions. Fluorescently tagged DNA sequences, which bind to chromosomal segments with a high degree of sequence complementarity, are used as probes, and a fluorescence-microscope is used to detect the presence or absence of the fluorescent signal. In addition, multi-colour FISH or spectral karyotyping (Speicher et al. 1996; Schröck et al. 1996) has been used in chromosome painting methods where each chromosome is labelled with a different fluorescent dye or combination of fluorescent dyes to scan a set of metaphase chromosomes for large-scale rearrangements and translocations.

Although chromosome painting allows rapid estimation of large chromosomal changes such as the presence or absence of specific variants, it is largely being used to detect large variants. Moreover, FISH has been used as a complement to sequencing approaches to determine the presence of SVs whose endpoints cannot be well defined by sequencing approaches (Kidd et al. 2008).

1.6.2 Microarray-based methods

Microarrays have been used to detect and genotype SVs (Pinkel et al. 1998; Iafrate et al. 2004; Locke et al. 2004; Sebat et al. 2004). These methods use hybridization between complementary DNA sequences as an indication for the presence or absence and quantity of chromosomal sequences in a high throughput fashion (Ylstra et al. 2006). Examples of microarray-based methods, notably array comparative genomic hybridization (array CGH) and SNP genotyping arrays, will be discussed in more detail in sections 1.6.2.1 and 1.6.2.2 respectively.

These microarray technologies provide no information on the location of duplicated copies and are not able to resolve breakpoints at a base-pair level. These technologies, however, offer a distinct advantage in terms of throughput and cost which make arrays a favored tool to discover SVs (Itsara et al. 2009; Li and Olivier 2013).

1.6.2.1 Array comparative genomic hybridization (array CGH)

In array CGH, fluorescently labelled samples hybridize to a microarray with a set of targets (typically long oligonucleotides) (Ylstra et al. 2006). The signal obtained from the level of fluorescence is a measure for the number of DNA segments in the query sample homologous to that target sequence. A reference or control sample is used to normalize the fluorescent signal of the target segments, which subsequently is used to identify potential gain and/or losses in a query genome. If only one sample is used, it is difficult to find whether it is because of the loss in

reference sample or it is a real gain in the query sample. Thus, the effect of the reference sample should be taken into consideration while interpreting results from array CGH (Park et al. 2010).

1.6.2.2 Single nucleotide polymorphism (SNP) arrays

The SNP microarray platforms are also based on hybridization and basically with little differences compared to aCGH platforms. In particular, probes on the array have been designed to identify specific single nucleotide variations between DNA sequences. This platform was originally designed to detect single nucleotide variations but subsequently was used to identify copy number variants as well. The abundance of SNP data from a large number of individuals, from efforts like the International HapMap Consortium, motivated additional studies on CNV detection (The International HapMap Project 2003). In this platform, the hybridization is performed on a single sample per microarray and log-transformed ratios are generated by clustering the intensities measured at each probe across many samples (Cooper et al. 2008). Patterns of SNPs provide evidence for different types of SVs, for example deletions appear as a run of null genotypes and do not fit the expected Mendelian inheritance from parent-child trios (Conrad et al. 2006; McCarroll and Altshuler 2007). Similarly, differences at the signal ratio between test and reference samples suggest the copy number of a particular segment in the query genome.

1.6.3 Sequence based approaches

Due to the advances in next generation sequencing (NGS) technology, DNA sequencing has become the dominant approach to detect SVs. NGS platforms (eg. Illumina HiSeq and Ion Torrent) produce large amounts of data with various read lengths and insert sizes. Most of the SV studies use available reference genomes to align or assemble these reads while searching for regions with discordant signatures or patterns. Such signatures of discordant mapping are then categorized

into different classes of SVs. Most of the current algorithms for SV discovery are modeled on computational methods that were first developed to analyze capillary sequencing reads and fully sequenced large-insert clones (Tuzun et al. 2005; Volik et al. 2003). There are four different strategies which utilize an available reference genome to align or assemble the sequencing reads and subsequently search for SVs, which I will discuss in more detail in sections 1.6.3.1 to 1.6.3.4.

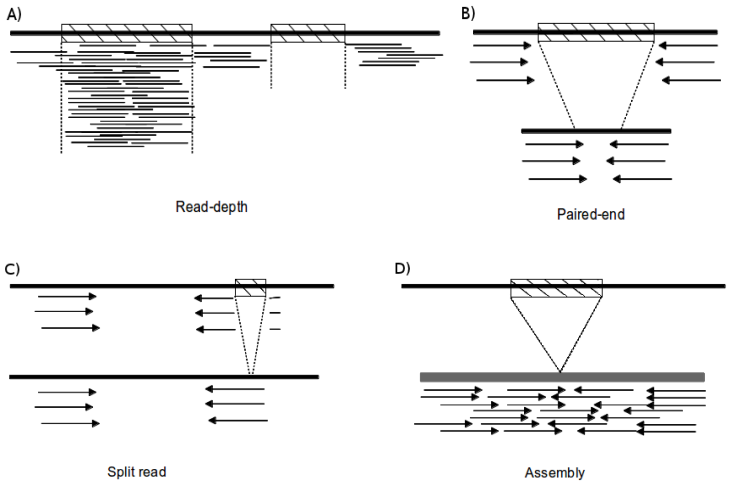


Figure 1.2 Different sequenced based approaches to detect SVs. A) Read depth method, B) Paired-end method, C) Split-read method, and D) assembly based approach.

1.6.3.1 Read depth approach

The availability of high coverage NGS data makes it possible to identify CNVs, based on the read depth of the sequence. The read depth (RD) approach assumes that the sequencing process is uniform and the number of reads mapping to a region follows a Poisson distribution. This approach also expects to have mapping depth comparable to the number of times a region appears in the donor genome. Hence, by looking at the divergence from the distribution of the read depth in the sequenced samples, deleted and duplicated regions are discovered (Bailey et al. 2002). Thus, a duplicated region will have a significantly higher read depth whereas

a deleted region will have a significantly lower read depth compared to the diploid regions of the same individual (Figure 1.2A).

The RD methods partition the reference into non-overlapping windows, and use reads mapped to each specific window as a proxy for the copy number of the window (Alkan et al. 2009; Chiang et al. 2009; Yoon et al. 2009; Sudmant et al. 2010; Bickhart et al. 2012; Esteve-Codina et al. 2013; Paudel et al. 2013). For example, Alkan et al., and Sudmant et al. used a set of known diploid regions in the human genome as control windows and calculated the average read depth for those regions. Similarly, in chapter 2, 3 and 4, we used 1:1 orthologous regions between distantly related species, in this case pig, cow, and human, as calibration/control region. The average read depth of those regions was used to calculate copy numbers (CNs) of other windows (Paudel et al. 2013). Finally, regions of gain and loss are extracted based on the copy number of each window. Other methods such as CNV-seq use a similar technique to call copy number but partition the reference genome in a sliding window (Xie and Tammi 2009). The RD approach using NGS data was first applied to define rearrangements in cancer genomes (Campbell et al. 2008; Chiang et al. 2009). It was later used to detect segmental duplications and generate copy number maps in human genomes (Alkan et al. 2009; Sudmant et al. 2010) followed by other mammalian genomes (Bickhart et al. 2012; Axelsson et al. 2013; Esteve-Codina et al. 2013; Paudel et al. 2013).

Although the RD approach is the only sequence-based method for accurate prediction of absolute CNs (Alkan et al. 2009; Sudmant et al. 2010; Bickhart et al. 2012; Esteve-Codina et al. 2013; Paudel et al. 2013), the breakpoint resolution is often poor. There are several limitations of the RD method. It is limited to the detection of CNVs, for example, SVs other than CNVs such as inversions, translocations, and novel insertions are not possible to assess with this approach. The high sequence similarity of repetitive regions in the genome is another

drawback of this approach. The challenge here is to deal with the adequate allocation of reads to those regions. To avoid this, the highly repetitive regions are masked prior to the alignment. Similar to the array CGH methods, it cannot provide the location of novel duplicated regions. A further weak point is the well-known bias of sequencing platforms towards the GC composition of the sequences. This bias needs to be properly addressed before calling CN. The final limitation of this approach is the sequence coverage, since the RD based methods depend on the signal-to-noise ratio, where the noise is primarily derived from the stochasticity of the RD, increased sequence coverage improves sensitivity, break point, and ultimately CN estimation.

1.6.3.2 Paired-end method

The paired-end method is an approach where paired-end reads are aligned against the reference genome and the discordantly aligned paired-end reads, in terms of orientation and position, are considered to detect SVs. In theory, using this approach most of the SVs can be identified. Paired-end reads that map too far from each other on the reference genome indicate that the region between mates is (partially) deleted, and those found being mapped too close indicate an insertion relative to the reference genome (Figure 1.2B). Similarly, the inconsistent orientation of the paired-end reads can represent inversions and tandem duplications (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008, 2010). Paired-end reads with pairs mapped on different chromosomes indicate the presence of translocations (Tuzun et al. 2005) whereas, novel insertions are discovered when only one end of paired-end reads cluster and the other ends do not align to the reference.

The accuracy of the predicted SVs using the pair-end method is highly dependent on the quality of reads, distribution of the insert size of read libraries, the mapping quality and the quality of the reference sequence.

Many different tools have been developed to detect SVs using the paired-end method. Some tools allow uniquely mapped reads only, like GASV (Sindi et al. 2009), PMer (Korbel et al. 2009), and Breakdancer (Chen et al. 2009a), whereas others allow multiple alignments of the paired-end reads to the reference genome such as VariationHunter (Hormozdiari et al. 2010). Two different strategies have been implemented to detect SVs using the paired-end method. The first is the cluster-based strategy implemented by PEMer, GASV, BreakdancerMAX and VariationHunter. In this approach, a fixed set of discordant mappings is selected that supports the same potential SV event, also called 'valid cluster' and predictions are made based on these clusters (Medvedev et al. 2009). A cluster should include a minimum of two paired-end reads to ensure the accuracy of the predication of breakpoints and the SV size (Medvedev et al. 2009). The second strategy, implemented by MoDILl(Lee et al. 2009), is called the model-based approach, which adopts a probability test to discover SVs by comparing the observed length distribution of paired-end reads at a particular location to the expected genome wide distribution of the insert length (Lee et al. 2009).

1.6.3.3 Split-read method

The split-read method allows to accurately detecting breakpoints of small insertions and large deletions at single base pair resolution. It only considers the paired-end reads for which one of the mates does not align or only partially aligns to the reference genome. The unaligned or partially aligned paired-end reads are re-aligned to the reference genome by splitting them into multiple fragments (Figure 1.2C). This realigning step therefore provides the precise start and end positions of the insertion or deletion event. This approach is not suitable to detect large insertion events in a genome.

The Pindel algorithm is the first algorithm to use the split-read approach to identify breakpoints of large deletions (1-10 kilobases) and small insertions (1-20 bases) from NGS data (Ye et al. 2009). It utilizes the paired-end reads approach to reduce the computational challenge of the locally gapped alignment of short sequences to the reference genome. For that, it first searches for the unaligned or partially aligned paired-end reads. The properly aligned reads of a pair are used as an anchor and a pattern growth approach is applied to determine the optimal alignment of split reads in minimum (the 5' end of the input reads) and maximum locations (the 3' end of the input reads).

1.6.3.4 Assembly approach

In the genome assembly approach, a query genome is assembled using short reads generated by NGS tools. In theory, de novo assembly of the query genome and a comparison to the reference genome can detect all forms of SVs present in the query genome. Recently, with the improvement of sequencing tools to generate longer and more accurate read fragments, this approach has emerged as a powerful method to detect SVs, however, available assembly algorithms are limitation in this approach. Most of the assembly based approaches use a combination of de novo assembly and local genome assembly to generate contigs (Figure 1.2D). These contigs are then compared to a reference genome to infer SVs. Some recent studies have implemented the local assembly approach to discover novel insertions in the human genome (Kidd et al. 2008, 2010). In these studies, researchers extracted the unmapped ends of paired-end reads. By using mapped ends of the paired-end reads as an anchor to the reference, the other reads were assembled to create larger fragments as contigs and referred to as novel insertions because they were absent in the reference genome (Hajirasouliha et al. 2010; Wang et al. 2011; Iqbal et al. 2012).

Comparing the *de novo* assembled genome to a very high quality reference genome can ideally yield all types of variations that occur in a query genome. However, due to the limitations of this approach such as the read length, sequence quality and computation power, the assembly approach has not been widely adopted yet. The de novo assembly algorithms such as EULER-USR (Chaisson et al. 2009), ABySS (Simpson et al. 2009), SOAPdenovo (Luo et al. 2012) and ALLPATHS-LG (Maccallum et al. 2009) use NGS data to assemble query genomes however, none of them are designed to detect SVs. Tools such as NovelSeq (Hajirasouliha et al. 2010), CREST (Wang et al. 2011), and Cortex (Iqbal et al. 2012) have been developed to utilize assembly based approaches to detect different forms of SVs.

1.7 Objectives and thesis outline

Few studies have used NGS data to understand the dynamics of SVs such as CNVs during the process of domestication and speciation (Bickhart et al. 2012; Axelsson et al. 2013; Sudmant et al. 2013). CNV studies in domesticated animals could not resolve the questions related to the impact of CNVs on the domestication process due to the lack of ancestral wild populations and proper samples from different biogeographic regions. Similarly, due to the absence of samples of evolutionarily closely related sub-species, no clear impact of CNVs on the process of speciation has been documented.

Pigs were domesticated several times, independently, from local wild populations in Asia and Europe (Larson et al. 2005; Megens et al. 2008). Due to the extensive selective pressures, differences in SVs in genomes between wild and domestic populations from the Eurasian region might reflect not only selection but also biogeography and domestication history of pigs. We have sequenced individuals of different populations of both wild and domestic pigs from Asia and Europe, which gave us a unique opportunity to understand the impact of different selection pressure on genomes. Similarly, we have sequenced different morphologically

defined species of the genus Sus from Island of South East Asia, i.e. Java, Borneo, Sulawesi, and The Philippines. These morphologically defined species are still capable of producing fertile offspring and the process of differentiation is ongoing (Blouch and Groves 1990), which gave us an opportunity to study the impact of SVs on the ongoing process of speciation. Hence, in this thesis, I will discuss the use of NGS data to improve our understanding of the role of SVs such as CNVs on the process of domestication, and their impact on the ongoing process of speciation. In chapter 2 of this thesis, I describe our study on the dynamics of CNVs in pigs in the context of adaptation and domestication. In chapter 3, I take the analysis to a different level and describe the role of CNVRs in speciation. CNVs were mapped in five closely related species of the genus Sus to provide detailed knowledge on the potential evolutionary role of CNVs between species. In chapter 4, I focus my study to understand the effect of selection and genetic drift on the copy number variation of one of the largest known gene family in mammalian genome, the olfactory receptor gene family, in pigs. In chapter 5, I describe the results of a study of other types of genomic variation in pigs such as short insertions and deletions and SNPs.

References

Agarwal S, Tafel AA, Kanaar R. 2006. DNA double-strand break repair and chromosome translocations. Mech Chromosom Translocat 5: 1075–1081.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.

Alvarez C, Akey J. 2011. Copy number variation in the domestic dog. Mamm Genome 1–20.

Antshel KM. 2007. Autistic spectrum disorders in velo-cardiofacial syndrome (22q11.2 deletion). J Autism Dev Disord 37: 1776–1786.

Arinami T. 2006. Analyses of the associations between the genes of 22q11 deletion syndrome and schizophrenia. J Hum Genet 51: 1037–1045.

Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495: 360–364.

Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun JY. 2010. Identification of copy number variations and common deletion polymorphisms in cattle. BMC Genomics 11: 232.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. Science 297: 1003 – 1007.

Bassett AS. 1998. 22q11 deletion syndrome in adults with schizophrenia. Am J Med Genet 81: 328–337.

Bassett AS. 2003. The schizophrenia phenotype in 22q11 deletion syndrome. Am J Psychiatry 160: 1580–1586.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22: 778 – 790.

Blouch RA, Groves CP. 1990. Naturally occurring suid hybrid in Java. Z Für Säugetierkunde 55: 270–275.

Brouwers N, Van Cauwenberghe C, Engelborghs S, Lambert J-C, Bettens K, Le Bastard N, Pasquier F, Montoya AG, Peeters K, Mattheijssens M, et al. 2012. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. Mol Psychiatry 17: 223–233.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40: 722 – 729.

Carvalho CMB, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH, et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat Genet 43: 1074–1081.

Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res 19: 336–346.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009a. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Meth 6: 677–681.

Chen WK, Swartz JD, Rush LJ, Alvarez CE. 2009b. Mapping DNA structural variation in dogs. Genome Res 39: 500 – 509.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 6: 99 – 103.

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. 2006. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38: 75–81.

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40: 1199–1203.

Critchlow SE, Jackson SP. 1998. DNA end-joining: from yeast to man. Trends Biochem Sci 23: 394–398.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell 149: 912–922.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17: 1266–1277.

Durkin K, Coppieters W, Drogemuller C, Ahariz N, Cambisano N, Druet T, Fasquelle C, Haile A, Horin P, Huang L, et al. 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature 482: 81–84.

Edelmann L, Pandita RK, Morrow BE. 1999. Low-Copy Repeats Mediate the Common 3-Mb Deletion in Patients with Velo-cardio-facial Syndrome. Am J Hum Genet 64: 1076–1086.

Elferink M, Vallee A, Jungerius A, Crooijmans R, Groenen M. 2008. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. BMC Genomics 9: 391.

Eliez S. 2007. Autism in children with 22Q11.2 deletion syndrome. J Am Acad Child Adolesc Psychiatry 46: 433–434.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320: 1629–1631.

Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silio L, Rodriguez M, Groenen M, Ramos-Onsins S, Perez-Enciso M. 2013. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics 14: 148.

Fadista J, Nygaard M, Holm LE, Thomsen B, Bendixen C. 2008. A snapshot of CNVs in the pig genome. PLoS One 3: e3916.

Fadista J, Thomsen B, Holm LE, Bendixen C. 2010. Copy number variation in the bovine genome. BMC Genomics 11: 284.

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SCL, de Smith A, Blakemore AIF, et al. 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39: 721–723.

Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, Casadio R, Russo V, Portolano B. 2010. An initial comparative map of copy number variations in the goat (Capra hircus) genome. BMC Genomics 11: 639.

Giuffra E, Tornsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, Anderson SI, Archibald AL, Andersson L. 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. Mamm Genome 13: 569 – 577.

Gokcumen O, Tischler V, Tica J, Zhu Q, Iskow RC, Lee E, Fritz MH-Y, Langdon A, Stütz AM, Pavlidis P, et al. 2013. Primate genome architecture influences structural variation mechanisms and functional consequences. Proc Natl Acad Sci 110: 15764–15769.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al. 2005. The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. Science 307: 1434 –1440.

Gu J, Lieber MR. 2008. Mechanistic flexibility as a conserved theme across 3 billion years of nonhomologous DNA end-joining. Genes Dev 22: 411–415.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. Pathogenetics 1: 4.

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC. 2010. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26: 1277–1283.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP. 2006. A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. Science 313: 101 –104.

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26: i350–i357.

Huffmeier U, Bergboer JGM, Becker T, Armour JA, Traupe H, Estivill X, Riveira-Munoz E, Mossner R, Reich K, Kurrat W, et al. 2010. Replication of LCE3C-LCE3B CNV as a Risk Factor for Psoriasis and Analysis of Interaction with Other Genetic Risk Factors. J Invest Dermatol 130: 979–984.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36: 949 – 951.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97–108.

Inoue K, Lupski JR. 2002. Molecular mechanisms for genomic disorders. Annu Rev Genomics Hum Genet 3: 199–242.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44: 226–232.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, et al. 2009. Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. Am J Hum Genet 84: 148–161.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56 – 64.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. Cell 143: 837–847.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, et al. 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol 10: e1001258.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. Spec Issue Mem John Maynard Smith Spec Issue Mem John Maynard Smith 239: 141–151.

Korbel J, Abyzov A, Mu X, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein M. 2009. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10: R23.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420 – 426.

Langer-Safer PR, Levine M, Ward DC. 1982. Immunological method for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci 79: 4381–4385.

Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E, et al. 2005. Worldwide Phylogeography of Wild Boar Reveals Multiple Centers of Pig Domestication. Science 307: 1618 –1621.

Lee JA, Carvalho CMB, Lupski JR. 2007. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. Cell 131: 1235–1247.

Lee S, Hormozdiari F, Alkan C, Brudno M. 2009. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods 6: 473 – 474.

Lieber MR. 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu Rev Biochem 79: 181–211.

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. Genome Res 39: 693 – 703.

Li W, Olivier M. 2013. Current analysis platforms and methods for detecting copy number variation. Physiol Genomics 45: 1–16.

Locke DP, Segraves R, Nicholls RD, Schwartz S, Pinkel D, Albertson DG, Eichler EE. 2004. BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. J Med Genet 41: 175–182.

Long M. 2001. Evolution of novel genes. Curr Opin Genet Dev 11: 673–680.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1: 18.

Lupski J. 2004. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. Genome Biol 5: 242.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14: 417–422.

Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, Ranade S, Shea TP, et al. 2009. ALLPATHS 2: small genomes

assembled accurately and with high continuity from short paired reads. Genome Biol 10: R103.

McCarroll SA, Altshuler DM. 2007. Copy-number variation and association studies of human disease. Nat Genet 39: S37 – S42.

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat Meth 6: S13–S20.

Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, Franke A, Malafosse A, Genton P, Thomas P, et al. 2010. Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. PLoS Genet 6: e1000962.

Megens HJ, Crooijmans R, San Cristobal M, Hui X, Li N, Groenen MA. 2008. Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genet Sel Evol 40: 103 – 128.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470: 59 – 65.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome Res 23: 749–761.

Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. 2009. The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Res 19: 491 –499.

Nobile C, Toffolatti L, Rizzi F, Simionati B, Nigro V, Cardazzo B, Patarnello T, Valle G, Danieli G. 2002. Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum Genet 110: 418–421.

Noor MAF, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci 98: 12084–12088.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, et al. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411: 603–606.

Otto SP, Yong P. 2002. The evolution of gene duplicates. In Advances in Genetics (ed. Jay C. Dunlap and C.-ting Wu), Vol. Volume 46 of, pp. 451–483, Academic Press.

Pang AWC, Migita O, MacDonald JR, Feuk L, Scherer SW. 2013. Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. Hum Mutat 34: 345–354.

Pankratz N, Kissell DK, Pauciulo MW, Halter CA, Rudolph A, Pfeiffer RF, Marder KS, Foroud T, Nichols WC, For the Parkinson Study Group–PROGENI Investigators. 2009. Parkin dosage mutations have greater pathogenicity in familial PD than simple sequence mutations. Neurology 73: 279–286.

Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D, et al. 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39: 830–832.

Park H, Kim J-I, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, et al. 2010. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42: 400–405.

Paudel Y, Madsen O, Megens H-J, Frantz L, Bosse M, Bastiaansen J, Crooijmans R, Groenen M. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14: 449.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256–1260.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W-L, Chen C, Zhai Y, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–211.

Ramayo-Caldas Y, Castello A, Pena RN, Alves E, Mercade A, Souza CA. 2010. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. BMC Genomics 11: 593.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD. 2006. Global variation in copy number in the human genome. Nature 444: 444 – 454.

Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, Wade CM, von Euler H, Gustafson U, Hedhammar A, et al. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat Genet 39: 1318 – 1320.

Schröck E, Manoir S du, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, et al. 1996. Multicolor Spectral Karyotyping of Human Chromosomes. Science 273: 494–497.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. 2004. Large-Scale Copy Number Polymorphism in the Human Genome. Science 305: 525 –528.

Shaw CJ, Bi W, Lupski JR. 2002. Genetic Proof of Unequal Meiotic Crossovers in Reciprocal Deletion and Duplication of 17p11.2. Am J Hum Genet 71: 1072–1081.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res 19: 1117–1123.

Sindi S, Helman E, Bashir A, Raphael BJ. 2009. A geometric approach for classification and comparison of structural variants. Bioinformatics 25: i222–i230.

Speicher MR, Ballard SG, Ward DC. 1996. Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat Genet 12: 368 – 375.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet 18: 74–82.

Sturtevant AH. 1920. Genetic studies on Drosophila simulans. Genetics 5: 488–500.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23: 1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 Genomes, et al. 2010. Diversity of human copy number variation and multicopy genes. Science 39: 641 – 646.

The International HapMap Project. 2003. The International HapMap Project. Nature 426: 789–796.

Toffolatti L, Cardazzo B, Nobile C, Danieli GA, Gualandi F, Muntoni F, Abbs S, Zanetti P, Angelini C, Ferlini A, et al. 2002. Investigating the Mechanism of Chromosomal Deletion: Characterization of 39 Deletion Breakpoints in Introns 47 and 48 of the Human Dystrophin Gene. Genomics 80: 523–530.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet 37: 727–732.

Van Gent DC, Hoeijmakers JHJ, Kanaar R. 2001. Chromosomal stability and the DNA double-stranded break connection. Nat Rev Genet 2: 196–206.

Volik S, Zhao S, Chin K, Brebner JH, Herndon DR, Tao Q, Kowbel D, Huang G, Lapuk A, Kuo W-L, et al. 2003. End-sequence profiling: Sequence-based analysis of aberrant genomes. Proc Natl Acad Sci 100: 7696–7701.

Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, et al. 2011. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Meth 8: 652–654.

Wiseman J. 1986. A history of the British pig. Ebenezer Baylis & Son Ltd. Worcester, UK.

Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, Vieaud A, Tixier-Boichard M, Rubin C-J, Imsland F, Hallböök F, et al. 2009. Copy Number Variation

in Intron 1 of SOX5 Causes the Pea-comb Phenotype in Chickens. PLoS Genet 5: e1000512.

Xie C, Tammi M. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10: 80.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871.

Ylstra B, van den IJssel P, Carvalho B, Brakenhoff RH, Meijer GA. 2006. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res 34: 445–450.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19: 1586–1592.

Yu K, Lieber MR. 2003. Nucleic acid structures and enzymes in the immunoglobulin class switch recombination mechanism. DNA Repair 2: 1163–1174.

2

Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication

Yogesh Paudel[1], Ole Madsen[1], Hendrik-Jan Megens[1], Laurent A. F. Frantz[1], Mirte Bosse[1], John W. M. Bastiaansen[1], Richard P. M. A. Crooijmans[1] and Martien A. M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, De Elst 1, 6700 AH, Wageningen, The Netherlands

BMC Genomics 2013, 14:449

Abstract

Copy number variable regions (CNVRs) can result in drastic phenotypic differences and may therefore be subject to selection during domestication. Studying copy number variation in relation to domestication is highly relevant in pigs because of their very rich natural and domestication history that resulted in many different phenotypes. To investigate the evolutionary dynamic of CNVRs, we applied read depth method on next generation sequence data from 16 individuals, comprising wild boars and domestic pigs from Europe and Asia.

We identified 3,118 CNVRs with an average size of 13 kilobases comprising a total of 39.2 megabases of the pig genome and 545 overlapping genes. Functional analyses revealed that CNVRs are enriched with genes related to sensory perception, neurological process, and response to stimulus, suggesting their contribution to adaptation in the wild and behavioral changes during domestication. Variations of copy number (CN) of antimicrobial related genes suggest an ongoing process of evolution of these genes to combat food-borne pathogens. Likewise, some genes related to the omnivorous lifestyle of pigs, like genes involved in detoxification, were observed to be CN variable. A small portion of CNVRs was unique to domestic pigs and may have been selected during domestication. The majority of CNVRs, however, is shared between wild and domesticated individuals, indicating that domestication had minor effect on the overall diversity of CNVRs. In addition, the excess of CNVRs in non-genic regions implies that a major part of these variations is likely to be (nearly) neutral. Comparison between different populations showed that larger populations have more CNVRs, highlighting that CNVRs are, like other genetic variation such as SNPs and microsatellites, reflecting demographic history rather than phenotypic diversity.

CNVRs in pigs are enriched for genes related to sensory perception, neurological process, and response to stimulus. The majority of CNVRs ascertained in domestic pigs are also variable in wild boars, suggesting that the domestication of the pig did not result in a change in CNVRs in domesticated pigs. The majority of variable regions were found to reflect demographic patterns rather than phenotypic.

Key words: structural variation, copy number variation, next generation sequencing data, read depth method

2.1 Introduction

Linking genotypic variation to phenotypic variation is one of the most challenging aspects of contemporary genome research. While several studies have found that single nucleotide polymorphisms (SNPs) can have drastic effects on phenotype (Hoekstra et al. 2006; Kijas et al. 2012), these types of variation are unlikely to solely explain the large phenotypic diversity found at the inter and intra specific level. Recent genomic studies have shown that variations, other than SNPs, such as structural variations (SVs) also play a prominent role in phenotypic evolution (Dennis et al. 2012).

Polymorphic SVs may lead to different copy number of specific genomic regions within a population. These regions are often called copy number variable regions (CNVRs) and can range from 50 bases up to several megabases (Mb) (Mills et al. 2011). CNVRs constitute roughly 5-12% of the human genome (Redon et al. 2006; Stankiewicz and Lupski 2010) and have been recognized as a source of phenotypic variation including susceptibility to specific diseases (Redon et al. 2006; Korbel et al. 2007; Kidd et al. 2008; Stankiewicz and Lupski 2010). Duplication of genic regions can also result in evolution of new genes and gene functions that can have a significant impact on phenotypes (Feuk et al. 2006; Freeman et al. 2006; Ibeagha-Awemu et al. 2008; Marques-Bonet et al. 2009; Zhang et al. 2009). For example, duplication of the *CCL3L1* gene can protect an individual against contracting HIV and developing AIDS (Gonzalez et al. 2005) and a partial duplication of the Slit-Robo Rho GTPhase-activating protein 2 gene (*SRGAP2*), some around 3 million years ago (mya), created a novel gene function associated with cognitive abilities in humans (Guerrier et al. 2009; Guo and Bao 2010; Dennis et al. 2012).

In domestic animals the best-known examples of traits that are affected by CNVRs pertain the animal exterior. For instance, a duplication of the agouti signaling

protein gene (*ASIP*) in sheep results in a different pigmentation (Norris and Whan 2008). The duplication of a set of fibroblast growth factor (*FGF*) genes in dogs leads to a characteristic dorsal hair ridge (Salmon Hillbertz et al. 2007). A copy number gain of the region containing the *KIT* gene causes the dominant white/patch coat phenotype observed in different European pig breeds (Pielberg et al. 2002, 2003). Thus, the association of CNVRs with distinct large effects in species that very recently have undergone strong phenotypic alteration, most notably domesticated animals in the past 10 thousand years, raises the question of how rapid phenotypic alteration may be related to (large) structural variation in genomes.

Sus scrofa (domesticated pigs and wild boars; family: Sudiae) diverged from other Sus species some 4 mya and started to spread, from Southeast Asia, into the rest of its currently natural occurrence across most of the Eurasia about 1.2 - 0.6 mya (Frantz LAF, unpublished observations). Such a large bio-geographic range will result in a wide range of local adaptation that, in part, may be related to CNVRs. Domestication can be seen as a long lasting genetic experiment (Megens and Groenen 2012), and in the case of pigs has been carried out on the same wild ancestral species independently at least once in Europe and once in Asia (Larson et al. 2005; Megens et al. 2008). Independent domestication implies independent breeding practices in Europe and Asia for several thousand years. Historical records revealed that breeding was more intensive in Asia than in Europe for centuries (White 2011). Different breeding regime led to intensive trading of breeds between Europe and Asia, especially at the onset of the industrial revolution when Europeans massively imported Asian breeds (White 2011; Groenen et al. 2012). Since the wild ancestor is still present throughout the entire natural range, among domesticated species, *Sus scrofa* provides a well suitable framework for studying effects of both adaptation and domestication on mammalian genome structure, such as CNVRs.

The recent completion of the porcine genome (Groenen et al. 2012) and the advent of high-throughput sequencing methods, now allow for a comprehensive screen of variation, including structural variation in the pig. Although several different methods e.g. SNP arrays and array CGH have been applied to screen for SVs, methods based on next generation sequencing (NGS) technology in general, and read depth (RD) based methods (Sudmant et al. 2010) in particular, revealed better performance in detecting CNVRs. The advantage of this approach is seen especially in and near highly duplicated genomic regions, such as segmental duplications (SDs) where most of the array based methods fail (McCarroll 2008; Alkan et al. 2011).

In this study the RD method was applied on NGS data of 16 *Sus scrofa* individuals, representing the diversity of both wild and domesticated pigs, firstly to detect SVs/CNVs in the pig genome and secondly to relate the evolution of SVs/CNVs to pig genomics features and to population/domestication histories.

2.2 Results

2.2.1 Data selection, copy number detection and definition of multi copy regions

In this study, 16 pigs were selected to cover a broad representation of pig diversity of both wild and domestic pigs. The selection of samples included three wild boars from Asia and three from Europe and five domesticated individuals from Asia and five from Europe (Table 2.1; Supplementary Table 2.1A). Whole genome re-sequenced data were obtained for the 16 samples with the average coverage per sample varying between 7x and 11x. Reads were aligned against the porcine reference genome (*Sus scrofa* build 10.2 (Groenen et al. 2012)) using mrsFAST (Hach et al. 2010). The RD method (Sudmant et al. 2010) was used to detect copy numbers (CNs) in the 16 pig individuals (see materials and methods for details).

From the estimated CN we defined regions of CN gains (termed multi copy regions (MCRs)) as regions ≥ 6 kilobases (Kb) and CN > 3. We detected 61,761 MCRs in the 16 individuals with individual numbers of MCRs ranging from 3,750 in an Asian domestic (AsD05) to 3,984 in a European wild boar (EuWB03). The average number of MCRs per individual was 3,860 covering 49.93 Mb (Table 2.1; Supplementary Table 2.1A). The size of the MCRs identified varied from the predefined minimum of 6 Kb to 122 Kb with an average size of 13 Kb. The majority of MCRs was found to be common in all 16 individuals. The number of MCRs that were found specific to single individual ranged from 0-12. Regions of CN loss were also identified, but we observed a positive correlation between sequence depth and regions of CN loss. With the used sequence coverage, this resulted in a considerable numbers of false positive CN losses (data not shown) and it was therefore decided to exclude CN losses from further analyses.

2.2.2 Copy number variable regions among pigs

CNVRs can be identified by comparing CN of the overlapping MCRs in different individuals. We identified 5,097 MCRs with their corresponding CN in the 16 individuals. The standard deviation (s.d.) of CN of each MCR was calculated and MCRs with a s.d. ≥0.7 among the 16 individuals were regarded as CNVRs. In total, 3,118 putative CNVRs were obtained with an average size of 13 Kb, comprising 39.72 Mb of the porcine genome (Supplementary Table 2.2A; See Figures 2.1; 2.3 and Supplementary figures 2.2 & 2.3 for examples of CNVRs). The CNVR density per chromosome varies from 0.85% on chromosome 18 to 2.29% on chromosome 2 (Supplementary Table 2.2B).

Table 2.1 Number and total size of multi copy regions in the 16 individuals1.

Region	Populations	Individual[1]	Sample	Read-depth[2]	Total MCR	Size (Mb)
Asia	Wild	AsWB01	Japanese WB	11	3764	48.9
		AsWB02	N. Chinese WB	10	3832	49.75
		AsWB03	S. Chinese WB	10.1	3953	51.23
	Domestic	AsD01	Meishan	9	3926	50.89
		AsD02	Meishan	9.1	3854	49.89
		AsD03	Xiang	8.1	3858	49.74
		AsD04	Xiang	8	3861	50.19
		AsD05	Jianquhai	10.5	3750	47.99
Europe	Wild	EuWB01	Dutch WB	9	3768	48.79
		EuWB02	Dutch WB	8	3816	49.2
		EuWB03	Italian WB	10	3984	51.47
	Domestic	EuD01	Large white	8	3909	50.59
		EuD02	Large white	8	3929	50.9
		EuD03	Landrace	8	3800	48.85
		EuD04	Duroc	7.1	3814	49.54
		EuD05	Pietrain	11	3943	51.14

[1]More details on individual information (Supplementary Table 2.1A)
[2]Average read-depth of the diploid region.

2.2.3 Experimental validation

We evaluated the accuracy of CNVRs prediction by quantitative real time-polymerase chain reaction (qPCR). Ten genic CNVRs, ten non-genic CNVRs and four diploid regions were randomly selected and tested using two distinct primer sets per locus. 23 of the 24 assays were successful and for those we found 100% agreement with our CNVRs predictions indicating a low false discovery call of CNVRs by the methodology and thresholds used in our analysis. Details of the qPCR primers can be found in Supplementary Table 2.4C. We also compared the

predicted CNVRs with known CNVRs. The region in chromosome 8 containing the *KIT* gene in the pig genome, which is known to be copy number variable between different European breeds confirms our results (Pielberg et al. 2002, 2003) (Figure 2.1).

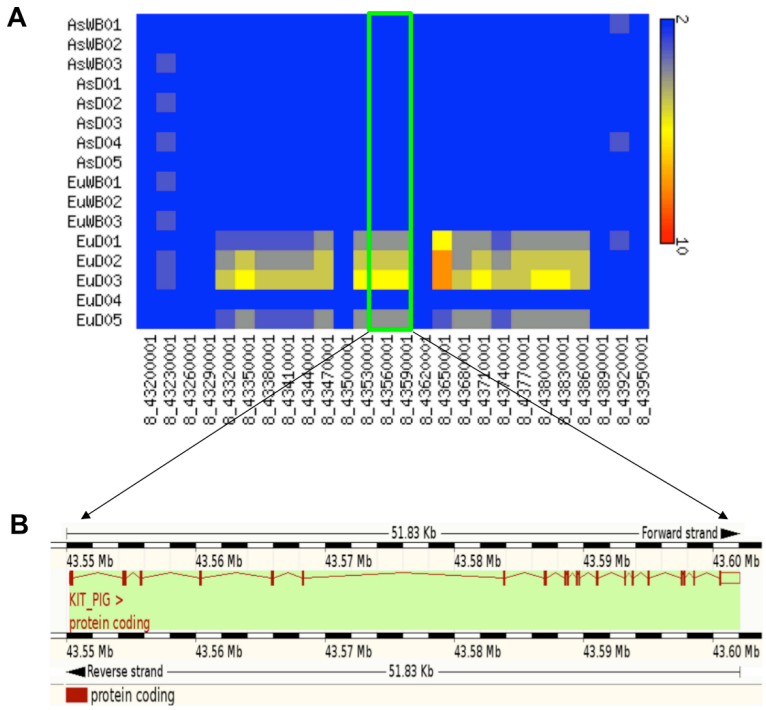


Figure 2.1 Region in chromosome 8 with the *KIT* gene.
The region in chromosome 8 with *KIT* gene (SSC8: 43,550,236-43,602,062), which is responsible for dominant white color in pigs shows an increase in the number of copies in the European domestic individuals.
A) Heatmap of the region containing the *KIT* gene. Blue color represents the diploid region where red color represents the region with copy number higher than 9.
B) Location of the *KIT* gene in the porcine genome (extracted from Ensembl browser).

2.3.4 Association of CNVRs with genomic features

Segmental duplications (SDs) (duplicated sequences larger than 1 Kb with more than 90% sequence similarity) act as promoter of CNVRs by facilitating non-allelic homologous recombination (Sharp et al. 2005; She et al. 2008). We compared the overlap between CNVRs with a list of 1,934 SDs previously identified in the pig

genome (Groenen et al. 2012). We found that approximately 27.5% of SDs (533 out of 1934) were overlapping within the 10 Kb flanking region of CNVRs. Both the CNVRs and SDs appear to be non-randomly distributed across the genome (Figure 2.2). Highly repetitive sequences such as retrotransposons were also investigated for their correlation with CNVRs. The frequencies of major retrotransposon families were calculated by counting the number of bases of these elements in the 10 Kb flanking regions of CNVRs and SD separately (Table 2.2). We observed significant enrichments of LINE-L1 (P <0.001, Fisher test), LTR-ERV1 (P <0.001, Fisher test) and satellite elements (P <0.001, Fisher test) near CNVRs and SDs (Table 2.2).

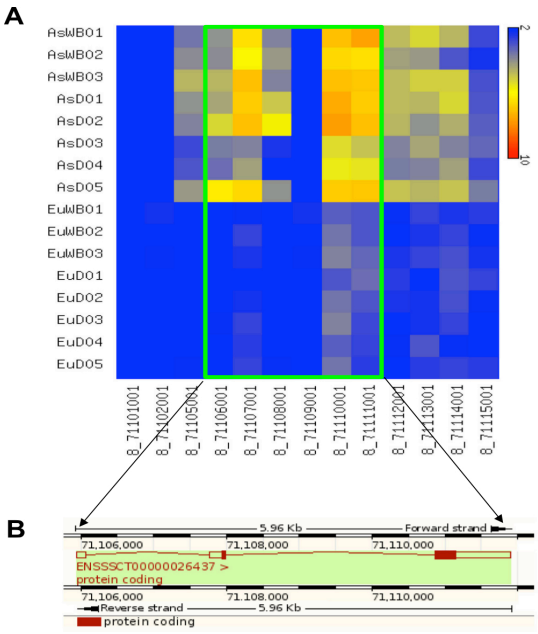


Figure 2.2 The *UGT2B10* gene in the porcine genome. The *UGT2B10* gene, which is involved in detoxification, shows increased copy number in the Asian individuals. A) Heatmap showing higher copies of *UGT2B10* (ENSSSCG00000026944; SSC8: 71,105,942-71,111,905) in Asian individuals (CN 5 to 9). B) Location of the *UGT2B10* in the porcine genome (extracted from Ensembl browser).

The guanine/cytosine (G/C) content of CNVRs and 10 Kb flanking region of CNVRs were assessed. Interestingly, it was observed that the G/C contents of CNVRs and 10 Kb flanking region of CNVRs are on average 1.5% and 1% lower than in the rest of the genome, respectively (Supplementary Table 2.2C).

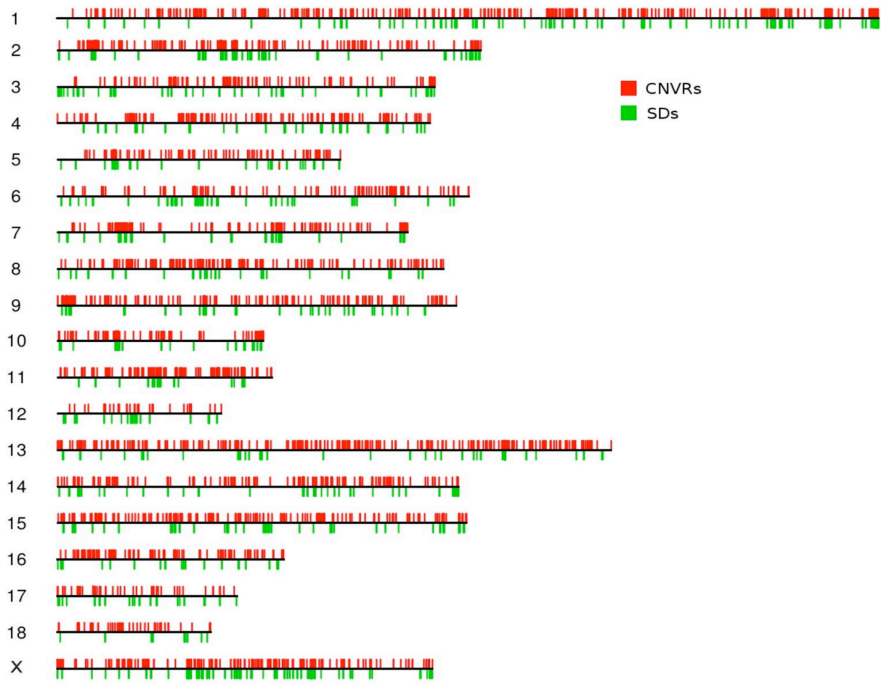


Figure 2.3 Distribution of CNVRs and SDs across the porcine genome. Black lines represent all 18 autosomes and the sex chromosome X. Red lines on the upper part of chromosomes indicate the 3,118 CNVRs and green lines on the lower part of chromosomes indicate 1,934 SDs.

2.3.5 Functional analysis of copy number polymorphic genes

Genes overlapping with CNVRs were extracted and potential functional roles associated with CNVRs were identified by analyzing them. Although partial duplication of a gene can lead to a functional new gene, the likelihood that a gene is functional intuitively decreases with the fraction of a gene that is duplicated. To limit the false discovery rate caused by the inclusion of a large fraction of non-functional gene duplicates, we only considered genes which are at least 70%

overlapping with a CNVR. Out of 21,627 genes annotated in the current genome build (*Sus scrofa* build10.2, Ensembl release 67 (Flicek et al. 2012; Groenen et al. 2012)), 575 protein-coding genes were found to overlap with the 454 CNVRs (14.56% of total CNVRs) (Supplementary Table 2.3A). A potential source of false positive calls are local high copy segments residing outside the gene exons resulting in CNVR calls without corresponding gene copy number variation. To avoid this type of false positives, the average depth of exon regions of the 575 genes, overlapping with a CNVR, were calculated (Supplementary Table 2.3A). Only genes with CN >2 in at least one individual and s.d. of ≥0.5 between 16 individuals were considered for further analysis. Of the 575 genes, 545 genes fulfilled this threshold (Supplementary Table 2.3B). Of the 11,629 one to one orthologous genes between human, cow and pig, only 25 were observed as multi copy genes including 10 olfactory receptor genes and genes like *KIT*, *BFAR*, *AHNAK* and *FLG2* (Supplementary Table 2.3C). Some of these genes only showed multiple copies in some of the individuals for example, *KIT* (Figure 2.1), whereas others showed high CN in all individuals like *FLG2* with CN ranging between 10-32.

The olfactory receptor gene family, one of the largest gene families in the porcine genome (Groenen et al. 2012; Nguyen et al. 2012), is over-represented with 353 out of 545 genes overlapping with CNVRs (Supplementary Table 2.3D). Genes involved in immune response, for instance *IFN* (Alpha-8, 11, 14; Delta-2), *IFNW1*, *IGK* (*V1D-43, V2-28, V8-61*), *IL1B* and *PG3I*, were often observed as variable in CN between individuals. Defense related genes *NPG3* and *PMAP23*, which are specific to porcine genome, were found to be variable in CN. In addition, genes involved in metabolism, *AMY1A, AMY2, AMY2A, AMY2B* and *BAAT*, and detoxification, *ABCG2, UGT2B10, UGT1A3, CYPA11, CYPA22, CYP4F3* and *CYP4X1*, are also present in the list of copy number variable genes.

Few CN variable genes were observed to be unique to a specific group of pigs; Asian domestics, Asian wild boars, European wild boars or European domestic. One example is the genomic region at chromosome 8, which contains the *UGT2B10* gene (SSC8: 71105001-71116000; Supplementary Table 2.3A) and was found to have a high CN specifically in Asian domestics and Asian wild boars (Figure 2.3). Similarly, *BTN1A1, CDK17, CDK20, F5, FLG2, MGAT4C, RALGDS* and *SUSD4* show variation in CN in all individuals but have comparatively high CN in the Asian domestic individuals.

Human orthologs of the porcine genes were used to analyze the functional enrichment of genes affected by CNVRs. Gene ontology (GO) enrichment analysis revealed that most of these genes were involved in biological processes regulating sensory perception of smell ($p<0.001$), signal transduction ($p<0.001$), neurological process ($p<0.001$) and metabolic process ($p<0.001$) (Supplementary Table 2.4A).

Table 2.2 Densities of repetitive element families in pig CNVRs and SDs.

Repeats	PigCNVRs[1]	PigSDs[2]	Other intervals[3]
Number of 10 Kb intervals	5304	2467	259660
LINE-L1	2872.95*	2852.95*	1368.88
LINE-L2	259.06	241.895	263.975
SINE-tRNA-Glu	1132.72	1133.05	1049.36
LTR-ERV1	248.19*	438.18*	148.055
LTR-ERVL-MaLR	170.467	183.131	159.755
SINE-MIR	193.498	209.735	233.435
DNA-hAT-Charlie	106.889	136.9616	111.46
Satellite	638.778*	576.016*	273.754

[1] Flanking 10 Kb regions of both end of CNVRs, all overlapping regions are merged.
[2] Flanking 10 Kb regions of SDs, all overlapping regions are merged
[3] Whole genome is divided into 10 Kb regions
* p-value (<0.001)

2.4 CNVRs between groups

The inclusion of pigs from the two independent domestications together with animals representing their wild ancestors enables preliminary investigation into whether the pattern of CNVRs was influenced by the process of domestication and/or the demographic history of pigs. For this particular comparison, to avoid any bias caused by sampling size, we included only 12 individuals, 3 from each of the 4 different groups based on their geographical origin/population (Asian wild, Asian domestic, European wild and European domestic) (Supplementary Table 2.1B). We compared the extent of overlap between the different groups and combination of the four groups and for each comparison, CNVRs were calculated separately (applying a threshold of ≥0.7 s.d. to call CNVRs) (Figure 2.4).

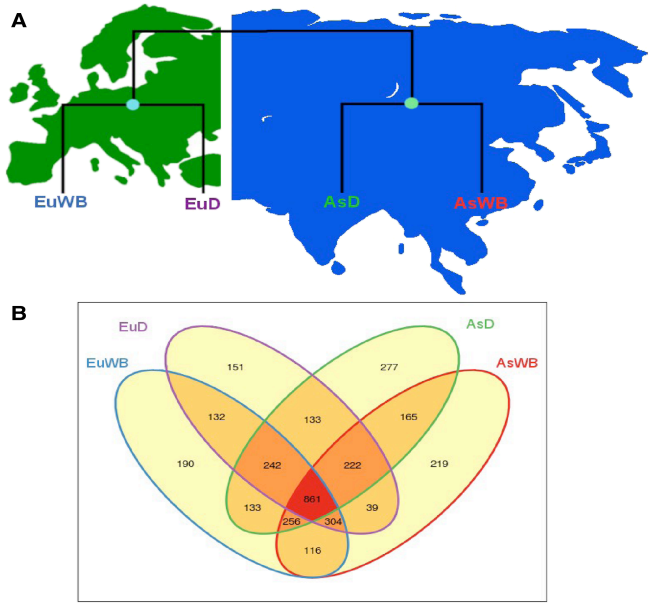


Figure 2.4 Pairwise comparison between different groups. A) Schematic representation of pigs across Eurasia. Two nodes show two independent domestication events. B) Shared CNVRs between different populations.

In all comparisons, it is evident that the large majority of CNVRs are shared among the different groups (Figure 2.4B). The Asian group (including both Asian wild and

Asian domestics) was found to have a higher CNVRs count (2,917) than the European group (2,779). Among the four groups, the Asian domestic group was found to have the largest number of CNVRs (2,289; of which 277 were group specific) with a ratio of 0.12 between shared and Asian domestic group specific CNVRs. The European domestic group was found to have the lowest number of CNVRs (2,084, 151 group specific) with a ratio of 0.07 between the shared and European domestic group specific CNVRs (Figure 2.4). Applying the same criterion as described above in the functional analyses, we extracted the genes overlapping with the CNVRs found in the comparative analyses. For each of the four groups we calculated the average cumulative count of genes and the s.d. of these overlapping genes (Supplementary Table 2.4B). It is notable that the number of genes situated in CNVRs seems to be higher in domesticated animals, both European and Asian, as compared to wild animals, but that the variation is lower in domesticated pigs compared to wild boars.

2.5 Discussion

Pigs have been important in agriculture and welfare for thousands of years. The recent completion of a high-quality draft genome of Sus scrofa (Groenen et al. 2012) enables the detailed investigation of a variety of genomics features. In this study, we used next generation sequence of 16 different wild as well as domestic pigs from Eurasia to generate a detailed map of CNVRs in the porcine genome.

2.5.1 CNVRs in pig genomes (compared to other mammalian genomes)

We applied the read depth methodology (Alkan et al. 2009; Sudmant et al. 2010; Bickhart et al. 2012) to estimate CNVRs. In total 3,118 CNVRs with an average size of 13 Kb were identified. Our result suggests that at least 1.5% (39.74 Mb) of the porcine genome can vary in CN of a size larger than 6 Kb, which is the minimum size we considered in this study. This figure is consistent with a recent study in cattle (Bickhart et al. 2012). It is likely that the actual count and size of variable regions in

the porcine genome is higher than our estimate. The stringent filtering criteria applied in our study, including a relatively high threshold of standard deviation to call a CNVR and exclusion of CN losses which were difficult to score with the sequence coverage currently available for the sampled individuals, likely inflated our false negative discovery rate. In addition, 100% validation of CNVRs tested by qPCR strengthens our confidence that our set of CNVRs is an underestimation rather than an overestimation.

Nevertheless, we estimated significantly more CNVRs than previously reported in pigs. Recently, two studies using array CGH inferred 259 CNVRs using 12 animals (Li et al. 2012) and 37 CNVRs on chromosomes 4, 7, 14 and 17 in a set of 12 samples. In addition, three other studies using the Porcine SNP60 genotypes inferred 49 CNVRs using 55 animals (Ramayo-Caldas et al. 2010), 382 CNVRs using 474 animals (Wang et al. 2012) and 565 CNVRs using 1693 pigs (Chen et al. 2012). The limitations faced by these studies, may be related to different factors such as, homogeneous sampling (only domestic pigs), low marker density, non-uniform distribution of SNPs along pig chromosomes and/or a lack of specially designed non-polymorphic probes which is necessary to identify CNVR with higher resolution (Ramos et al. 2009). Here, the RD method based on next-generation sequencing, using 16 different wild as well as domestic pigs from Eurasia, resulted in a better resolution and higher confidence to call CNVRs. Thus, most of the CNVRs discovered in this study are novel relative to the previous studies and represents the largest catalog of porcine specific CNVRs to date.

2.5.2 Association of CNVRs with genomic features

Previous studies suggested that repetitive elements play an important role in the formation of CNVRs and SDs (Cahan et al. 2009). Frequent breakage of DNA in and around the repeat regions could initiate non-allelic homologous recombination

(NAHR) and result in CNVRs (Hastings et al. 2009). The enrichment of the repetitive elements LINE-L1, LTR-ERV1 and satellite elements at the boundaries of CNVRs and SDs in the porcine genome (2.2), suggests that these families of repeat elements indeed facilitate the formation of CNVRs and SDs in the porcine genome. This is in accordance with the observation made by Giuffra et al. (2002), who has reported an association of LINE-L1 and the duplication of the region containing the KIT gene in the porcine genome (Giuffra et al. 2002). Similarly, the slightly lower G/C content (1.5%) of CNVRs in the porcine genome suggests that the porcine CNVRs are likely to coincide with the gene-poor regions, which is consistent to the observation made in the human genome (Yim et al. 2010).

2.5.3 Copy number polymorphic genes

In total, we found 545 genes overlapping with CNVRs representing a valuable resource for future studies on the relation between CNV genes and phenotype variation. Functional enrichment analysis suggests that genes involved in sensory perception of smell, signal transduction, neurological system process and metabolism are affected by the CNVRs. The enrichment of CNVRs involved in the sensory related genes is consistent to the general behavior of pigs, showing strong reliance on their sense of smell in various behavioral contexts. Collectively, this data might assist future studies on some of the genetic variation influencing morphological, behavioral and physiological traits in pigs.

Genes involved in immune response such as interferon (IFN), cytochrome P450 (CYP), are usually fast evolving due to their importance for the organism to respond rapid changes in the environment. Our results show that these type of genes are often found to be CN variable in pigs. For example, members of interferon (IFN) gene families, involved in defense against viral infections, and CYP genes, which are responsible for detoxification and drug metabolism, were found to be CN variable. Olfactory receptor (OR) represents another gene family that is over-represented in

our list of CN variable genes. Sus scrofa have the largest repertoire of functional OR genes in mammals (from mammals whose genome has been sequenced to date) (Nguyen et al. 2012), likely related to the strong dependence on their sense of smell for foraging (Groenen et al. 2012). Nearly one-third of the 1301 porcine OR genes are found as copy number variable in pigs. These findings suggest that the wide variety of environment faced by pigs around the world resulted in CNVs.

Among defense related copy number variable genes, *NPG3* (from 4 to 23 copies) and *PMAP23* (from 2 to 13 copies) are cathelicidin related porcine specific genes. NPG3 is responsible for microbicidal activity against Escherichia coli, Listeria monocytogenes and Candida albicans in vitro (Kokryakov et al. 1993) whereas *PMAP23* exerts antimicrobial activity against both gram-positive and gram-negative bacteria in vitro (Zanetti et al. 1994). In addition, *CAMP* (from 3 to 16 copies), another cathelicidin related gene present in the list of copy number variable genes. The observed variation in copy number of cathelicidin related genes suggests an ongoing process of evolution of this gene-family in porcine genome to combat food-borne pathogens.

In humans, copy number of amylase genes, especially *AMY1*, shows high variation between populations (from 2 to 15 copies). High copy number of *AMY1* allows more efficient breakdown of starch (Perry et al. 2007). Unlike in humans, pigs have a universally high number of copies (from 8 to 21 copies) of amylases (*AMY1*, *AMY2A*, *AMY2B*) between all individuals, suggesting universal importance of amylases for digesting starch-rich food in this omnivorous species.

Genes such as *BTN1A1* and *F5* are found to be involved in the regulation of milk lipid droplets (Ogg et al. 2004) and preterm delivery in human (Hao et al. 2004), respectively. Interestingly we found that these genes had variable numbers of

copies in different pig breeds. Specifically, Asian breeds have typically a higher number of copies of these genes. In the pig breeding industry, Asian breeds are famous for being highly prolific; with some breeds typically bearing more than 15 live young per litter. These results suggest that these genes have been important in the selection process for highly fertile breeds in Asia. It is notable that some of these fertility genes have high CN in some European breeds (especially Large whites). Recent studies shown that this particular breed has been extensively admixed with Chinese pigs in order to improve fertility traits during the industrial revolution (White 2011; Groenen et al. 2012). Thus, this pattern could also be the result of this well-known admixture.

Some members of the uridine diphosphate glucuronosyl transferases (UGTs) superfamily are found variable in copy number. UGTs are part of important metabolic pathways responsible for the detoxification and elimination of many different endobiotics and xenobiotics (Miners et al. 2006). The *UGT2B10* gene, which is one of the most important genes involved in N-glucuronidation of nicotine, has a higher copy number in Asian individuals (from 5 to 9 copies) than the European individuals (3 copies). The elevated copy number may be related to the ability to detoxifying specific plant secondary metabolites. Although, at present there is no data on wild boar feeding habits in relation to floristic differences between East and West Eurasia, our finding can direct future ecological studies on that subject.

2.5.4 Demography shape CNVR diversity

Regardless of their geographic origin, different pig populations have undergone different selective pressure. Important events were the foundation of modern pig breeds starting around 200 years ago during the industrial revolution, and more recently, the development of modern breeding practices in the past five decade in different parts of Asia and Europe.

The association of CNVRs with distinct phenotypic effect and different selective regimes in Europe and Asia, suggest that differences in structural variation between wild and domestic pigs as well as Asian and European populations, could reflect domestication history. By including different pigs from the two independent domestications together with individuals representative of their wild ancestors, enabled a first preliminary insight into the change in pattern of CNVRs influenced by the process of domestication and/or the natural demographic history of pigs.

To investigate the importance that CNVRs may have had on phenotypic diversification in breeds, we compared the amount of CNVRs in domesticated and wild individuals. We found more CNVRs in domesticated animals (2,915) than in wild boars (2,879). Moreover, our results showed that CNVR counts were also higher in Asian pigs (combined wild and domestic) (2,967) than in European pigs (2,779) (combined wild and domestic) (Figure 2.4), which is consistent with a large effective population size and diverse origin of Asian pigs (Megens et al. 2008; Groenen et al. 2012).

A recent study based on SNPs identified a similar pattern not only between breeds and wild but also between Asian and European pigs (Groenen et al. 2012). Thus, CN seems to be more variable in larger populations, following the similar patterns as other types of variation such as SNPs (Groenen et al. 2012) and microsatellites (Megens et al. 2008). This indicates that the general pattern of CNV is more reflecting demography rather than phenotypic diversity. Having large fractions of common CNVRs between different groups and excess of CNVRs (2,664; 85.43%) in non-genic regions suggest that a major part of these variations are likely to be neutral or nearly neutral. This further supports their reflection on demography rather than phenotypic diversity. These results are of importance as they show that

intensive artificial selection did not affect the overall diversity of CNVRs in domestic pigs and do not appear to be the major source of the large phenotypic diversity observed in domestic pigs.

2.6 Conclusion

We identified 3,118 CNVRs with an average size of 13 Kb comprising 39.2 Mb of the porcine genome, which represents the largest source of genetic variation identified in the porcine genome to date. The inferred CNV regions include 545 genes providing an important resource for future analyses on phenotypic variation in pigs. Functional analyses revealed CNVRs enriched for genes related to sensory perception, neurological process, and response to stimulus in specific breeds or wild population. Comparison between wild and domestic groups shows that, beside few exceptions, domestication did not lead to a change in CNVRs among breeds. Moreover, we found that most CNVRs ascertained in domestics were also variable in wild boars. This result suggests that the majority of CNVRs were already segregating among wild boars before domestication. Furthermore, while we identify few CNVRs that may be under selection during domestication and may lead to phenotypic differences, the majority of variable regions were found to reflect demographic pattern rather than selective regimes. Our study represent a comprehensive analysis of CNV in both domestic and wild pigs and provides valuable insight in the evolutionary dynamics of copy number variation, in the context of adaptation and domestication.

2.7 Materials and Methods

2.7.1 Database

In total 16 different individuals originated from 13 populations of Sus scrofa were sequenced at different sequencing centers using the Illumina HiSeq platform. The libraries are 100 bases pair-end reads with coverage per animal ranging between 7 – 11x. The sampled pigs comprised of three European wild boars (2- Dutch and 1-

Italian), five European domestics (2- Large whites and 1- from each Landrace, Duroc and Pietrain), three Asian wild boars (1- North Chinese, 1- South Chinese and 1- Japanese) and five Asian domestics (2- Meishan, 2- Xiang and 1- Jianquhai) (Table 2.1; Supplementary Table 2.1A). DNA samples were obtained from blood samples collected by veterinarians according to national legislation or from tissue samples from animals obtained from the slaughterhouse or in the case of wild boar from animals culled within wildlife management programs.

2.7.2 Sequence alignment and copy number estimation

Copy number of regions in the genomes of all the 16 individuals was detected by the read depth (RD) method (Alkan et al. 2009; Sudmant et al. 2010), where the number of copies present is inferred from sequence depth of whole genome sequence data. To calculate the average read depth from those libraries, reads were aligned to the available pig reference genome (Sus scrofa build 10.2) using mrsFAST v2.3.0.2 ("Micro-read (substitutions only) fast alignment and search tool" (Hach et al. 2010)) with an edit distance of at most 7. mrsFAST is a memory efficient and fast software, which reports all possible mapping locations (not only the best, unique or first mapping locations as several other softwares), which is essential in order to detect multi-copy regions using read depth method. Because the RD methods do not take paired end information into consideration, all the paired end libraries were treated as single end libraries.

Highly repeated elements are the main source of noise for the RD method. The porcine genome consists of more than 40 percent of highly repeated elements and most of these repeated elements are long/short interspersed nuclear elements (LINEs/SINEs), long terminal repeats retro-transposons (LTRs) and satellites (Groenen et al. 2012). To avoid noise from these repeated elements, a repeat masked reference genome was used. Repeat masked information was obtained from NCBI

(ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Sus_scrofa/Sscrofa10.2/Primary_Assembly/assembled_chromosomes/FASTA/) and merged with the repeat masked information used in Groenen et al. (2012) (Groenen et al. 2012). Calculation of read depth across the whole genome was done with the help of SAMtools v0.1.12a (Li et al. 2009). Average read depth for each 1 Kb non-overlapping bin was calculated across the genome.

RD method uses read depth information of diploid region to infer copy number of each 1 Kb non-overlapping bin present in the genome. No prior information regarding diploid regions in the porcine genome was available. We therefore used 1:1 orthologous genic regions between human, cow and pig as diploid region in the first stage to identify CN of each bin present in the genome (Supplementary Figure 2.1). Since, coding regions are known to have a higher G/C content than an average region of a genome (Högstrand and Böhme 1999; Galtier et al. 2001)
this procedure may introduce a G/C biased read depth. To reduce possible G/C bias caused by the 1:1 orthologous regions, all diploid regions predicted from 1:1 orthologous regions in the first stage were subsequently used to calculate the average diploid read depth of the porcine genome (Supplementary Figure 2.1).

Next generation sequencing methods has been shown to introduce a bias in the coverage in regions of high or low G/C. One of the major reason for GC bias coverage in Illumina sequences originates from the polymerase chain reaction (PCR) amplification step during library preparation as well as for cluster amplification on the Illumina flowcell (Oyola et al. 2012). This issue is similar for any sequencing technology that relies on PCR amplification (Quail et al. 2012). To correct for this bias we calculated G/C intervals correction factors as described by Sudmant et al. (2010) (Sudmant et al. 2010). These factors were used to correct read depth of each 1 Kb bin across the genome. CN of each 1 Kb non-overlapping bins were then estimated based on the G/C corrected read depth. Since the

samples include both male and female individuals, copy number of male X chromosomes were corrected by multiplying the read depth by 2 (outside the pseudo-autosomal regions) to make them comparable with female individuals.

2.7.3 Prediction of MCRs and defining CNVRs

All the 1 Kb bins with minimum CN of 1 were extracted from all 16 individuals and bins with CN >3 were chained to form multi copy regions (MCRs). The same MCRs might be assigned with different boundaries in different individuals due to technical and/or biological reason and therefore all the MCRs from all individuals were extracted merged and the CN of those regions for all 16 individuals were compared. Copy number variable regions were identified based on the standard deviation of the CN of MCRs in all 16 individuals. Hence, CNVR status was assigned to those regions, which were variable (s.d. ≥0.7) in CN across all 16 individuals.

2.7.4 Gene identification and Gene Ontology

All the annotated porcine genes from Sus scrofa build 10.2, Ensembl release 67, were extracted using Biomart (Haider et al. 2009) and genes which were overlapping with the CNVRs (≥70% overlap) were identified. To reduce false calls of particular genes as being multi copy genes, exons of genes overlapping with CNVRs were tested for average CN. GC correction on the read depth of all exons was performed using the correction factors obtained previously for the whole genome. All the genes with an average depth in exon regions >2 were kept in the list of genes affected by CNVRs for further analysis. Not all pig genes have associated gene names, thus the genes without gene names were blasted against the human Refseq mRNAs and human reference protein sequences (blastn and blastp respectively) and the best human hit was assigned as gene name. Human orthologs of porcine genes were used to perform gene ontology analysis. BinGO v2.44 (Maere et al. 2005) a plugin of Cytoscape v2.8.3 (Shannon et al. 2003) was used to identify enriched GO terms using human gene annotation as background.

Hypergeometric test was used to assess the significance of the enriched terms and Benjamini and Hochberg correction was implemented for multiple comparisons.

2.7.5 Comparison between different groups

For the group comparison, we formed groups based on their geographical location and population type (Asian wild, Asian domestic, European wild and European domestic). To make all the groups comparable with each other, we took 12 instead of all 16 individuals i.e. three pigs per group (Supplementary Table 2.1B). CNVRs for all groups were generated based on the similar approach we used before but instead of all 16 individuals, we compared only individuals present in the particular group.

2.7.6 qPCR Validation

Primer3 webtool http://frodo.wi.mit.edu/primer3/ was used to design primers for qPCR validation. Amplicon length was limited between (50 bp – 100 bp) and regions with GC percentage between 30% and 60% were included, while avoiding runs of identical nucleotides. All other settings were left at their default. Details of the qPCR primers can be found in Supplementary Table 2.4C. qPCR experiments were conducted using MESA Blue qPCR MasterMix Plus for SYBR Assay Low ROX from Eurogentec, this 2x reaction buffer was used in a total reaction volume of 12.5µl. All reactions were amplified on 7500 Real Time PCR system (Applied Biosystems group). The copy number differences were determined by using a standard ΔCt method that compares the mean Ct value of the target CNV fragments, determined from different input concentrations, compared to the mean Ct value of a known diploid reference.

2.7.7 Competing interests

The authors declare that they do not have any competing interests.

2.8 Author's contributions

OM, YP, H-JM, MAMG conceived and designed the experiments. YP, OM performed the experiments and analyzed the data. MAMG RPMAC contributed reagents/materials/analysis tools. YP wrote the manuscript. YP, OM designed and improved pipeline for CNV detection. OM MAMG H-JM LAFF MB JWMB RPMAC discussed and improved manuscript. All authors read and approved the final manuscript.

2.9 Acknowledgements

This work was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC Grant agreement no 249894 (SelSweep project). We would like to thank the Swine Genome Consortium for the reference genome build 10.2. We thank Prof. Dr. Ning Li, State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, China, for providing us DNA samples from Asian wild boars. We thank Bert Dibbits, Animal Breeding and Genomics Centre, Wageningen University for the qPCR validation and Dr. Anna Esteve Codina, Centre for Research in Agricultural Genomics (CRAG), Universitat Autonoma Barcelona and Dr. Roeland van Ham, Keygene N.V. for discussion.

2.10 Additional information

Published version of this chapter can be found here: http://www.biomedcentral.com/1471-2164/14/449

Supplementary files and tables can be downloaded from this link: http://www.biomedcentral.com/1471-2164/14/449/additional

References

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363–376.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome res 22: 778 – 790.

Cahan P, Li Y, Izumi M, Graubert TA. 2009. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nat Genet 41: 430–437.

Chen C, Qiao R, Wei R, Guo Y, Ai H, Ma J, Ren J, Huang L. 2012. A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. BMC Genomics 13: 733.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell 149: 912–922.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7: 85 – 97.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. Nucleic Acids Research 40: D84 –D90.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al. 2006. Copy number variation: New insights in genome diversity. Genome Research 16: 949 –961.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics 159: 907 –911.

Giuffra E, Tornsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, Anderson SI, Archibald AL, Andersson L. 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. Mamm Genome 13: 569 – 577.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al. 2005. The Influence of CCL3L1

Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. Science 307: 1434 –1440.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Guerrier S, Coutinho-Budd J, Sassa T, Gresset A, Jordan NV, Chen K, Jin W-L, Frost A, Polleux F. 2009. The F-BAR Domain of srGAP2 Induces Membrane Protrusions Required for Neuronal Migration and Morphogenesis. Cell 138: 990–1004.

Guo S, Bao S. 2010. srGAP2 Arginine Methylation Regulates Cell Migration and Cell Spreading through Promoting Dimerization. Journal of Biological Chemistry 285: 35133 –35141.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Meth 7: 576–577.

Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A. 2009. BioMart Central Portal—unified access to biological data. Nucleic Acids Research 37: W23 –W27.

Hao K, Wang X, Niu T, Xu X, Li A, Chang W, Wang L, Li G, Laird N, Xu X. 2004. A candidate gene association study on preterm delivery: application of high-throughput genotyping technology and advanced statistical methods. Human Molecular Genetics 13: 683 –691.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10: 551–564.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP. 2006. A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. Science 313: 101 –104.

Högstrand K, Böhme J. 1999. Gene conversion of major histocompatibility complex genes is associated with CpG-rich regions. Immunogenetics 49: 446–455.

Ibeagha-Awemu E, Kgwatalala P, Ibeagha A, Zhao X. 2008. A critical analysis of disease-associated DNA polymorphisms in the genes of cattle, goat, sheep, and pig. Mammalian Genome 19: 226–245.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, et al. 2012. Genome-Wide Analysis of the

World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol 10: e1001258.

Kokryakov VN, Harwig SSL, Panyutich EA, Shevchenko AA, Aleshina GM, Shamova OV, Korneva HA, Lehrer RI. 1993. Protegrins: leukocyte antimicrobial peptides that combine features of corticostatic defensins and tachyplesins. FEBS Letters 327: 231–236.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318: 420 –426.

Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E, et al. 2005. Worldwide Phylogeography of Wild Boar Reveals Multiple Centers of Pig Domestication. Science 307: 1618 –1621.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078 –2079.

Li Y, Mei S, Zhang X, Peng X, Liu G, Tao H, Wu H, Jiang S, Xiong Y, Li F. 2012. Identification of genome-wide copy number variations among diverse pig breeds by array CGH. BMC Genomics 13: 725.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21: 3448 –3449.

Marques-Bonet T, Girirajan S, Eichler EE. 2009. The origins and impact of primate segmental duplications. Trends in Genetics 25: 443–454.

McCarroll SA. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genet 40: 1166–1174.

Megens H-J, Groenen MAM. 2012. Domesticated species form a treasure-trove for molecular characterization of Mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. Heredity 109: 1–3.

Megens HJ, Crooijmans R, San Cristobal M, Hui X, Li N, Groenen MA. 2008. Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genet Sel Evol 40: 103 – 128.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470: 59–65.

Miners JO, Knights KM, Houston JB, Mackenzie PI. 2006. In vitro–in vivo correlation for drugs and other compounds eliminated by glucuronidation in humans: Pitfalls and promises. Biochemical Pharmacology 71: 1531–1539.

Nguyen D, Lee K, Choi H, Choi M, Le M, Song N, Kim J-H, Seo H, Oh J-W, Lee K, et al. 2012. The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. BMC Genomics 13: 584.

Norris BJ, Whan VA. 2008. A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. Genome Research 18: 1282 –1293.

Ogg SL, Weldon AK, Dobbie L, Smith AJH, Mather IH. 2004. Expression of butyrophilin (Btn1a1) in lactating mammary gland is essential for the regulated secretion of milk–lipid droplets. Proceedings of the National Academy of Sciences of the United States of America 101: 10084 –10089.

Oyola S, Otto T, Gu Y, Maslen G, Manske M, Campino S, Turner D, MacInnis B, Kwiatkowski D, Swerdlow H, et al. 2012. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics 13: 1.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256–1260.

Pielberg G, Day AE, Plastow GS, Andersson L. 2003. A Sensitive Method for Detecting Variation in Copy Numbers of Duplicated Genes. Genome Research 13: 2171 –2177.

Pielberg G, Olsson C, Syvänen A-C, Andersson L. 2002. Unexpectedly High Allelic Diversity at the KIT Locus Causing Dominant White Color in the Domestic Pig. Genetics 160: 305 –311.

Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, Bertoni A, Swerdlow H, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Ramayo-Caldas Y, Castello A, Pena RN, Alves E, Mercade A, Souza CA. 2010. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. BMC Genomics 11: 593.

Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, et al. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One 4: e6524.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. Nature 39: 444 – 454.

Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, Wade CM, von Euler H, Gustafson U, Hedhammar A, et al. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat Genet 39: 1318 – 1320.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13: 2498 –2504.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al. 2005. Segmental Duplications and Copy-Number Variation in the Human Genome. The American Journal of Human Genetics 77: 78–88.

She X, Cheng Z, Zollner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. Nat Genet 39: 909 – 914.

Stankiewicz P, Lupski JR. 2010. Structural Variation in the Human Genome and its Role in Disease. Annu Rev Med 61: 437–455.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 Genomes, et al. 2010. Diversity of human copy number variation and multicopy genes. Science 39: 641 – 646.

Wang J, Jiang J, Fu W, Jiang L, Ding X, Liu J-F, Zhang Q. 2012. A genome-wide detection of copy number variations using SNP genotyping arrays in swine. BMC Genomics 13: 273.

White S. 2011. From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. Environmental History 16: 94 –120.

Yim S-H, Kim T-M, Hu H-J, Kim J-H, Kim B-J, Lee J-Y, Han B-G, Shin S-H, Jung S-H, Chung Y-J. 2010. Copy number variations in East-Asian population and their evolutionary and functional implications. Human Molecular Genetics 19: 1001–1008.

Zanetti M, Storici P, Tossi A, Scocchi M, Gennaro R. 1994. Molecular cloning and chemical synthesis of a novel antibacterial peptide derived from pig myeloid cells. Journal of Biological Chemistry 269: 7855 –7858.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy Number Variation in Human Health, Disease, and Evolution. Annu Rev Genom Human Genet 10: 451–481.

3

Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors

Yogesh Paudel[1], Ole Madsen[1], Hendrik-Jan Megens[1], Laurent A. F. Frantz[1], Mirte Bosse[1], Richard P. M. A. Crooijmans[1] and Martien A. M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, De Elst 1, 6700 AH, Wageningen, The Netherlands

Submitted 2014

Abstract

Unraveling the genetic mechanisms associated with reduced gene flow between genetically differentiated populations is key to understand speciation. Different types of structural variations (SVs) have been found as a source of genetic diversity in a wide range of species. Previous studies provided detailed knowledge on the potential evolutionary role of SVs, especially copy number variations (CNVs), between well diverged species of e.g. primates. However, our understanding of their significance during ongoing speciation processes is limited due to the lack of CNV data from closely related species. The genus Sus (pig and its close relatives) which started to diverge ~4 Mya presents an excellent model for studying the role of CNVs during ongoing speciation.

In this study, we identified 1408 CNV regions (CNVRs) across the genus Sus. These CNVRs encompass 624 genes and were found to evolve ~2.5 times faster than single nucleotide polymorphisms (SNPs). The majority of these copy number variable genes are olfactory receptors (ORs) known to play a prominent role in food foraging and mate recognition in Sus. Phylogenetic analyses, including novel Bayesian analysis, based on CNVRs that overlap ORs retain the well-accepted topology of the genus Sus whereas CNVRs overlapping genes other than ORs show evidence for random drift and/or admixture.

We hypothesize that inter-specific variation in copy number of ORs provided the means for rapid adaptation to different environments during the diversification of the genus Sus in the Pliocene. Furthermore, these regions might have acted as barriers preventing massive gene flow between these species during the multiple hybridization events that took place later in the Pleistocene suggesting a possible prominent role of ORs in the ongoing Sus speciation.

Key words: speciation, structural variation, copy number variation, next generation sequencing data, read depth method

3.1 Introduction

The process of speciation is one of the major evolutionary drivers of the diversity of life on earth. Understanding the process by which populations diversify leading, ultimately, to speciation has been one of the major focuses of evolutionary biologists for decades (Mayr 1963; Mallet 1995; Coyne and Orr 2004). Two major models of speciation have been put forward. The first model, also known as allopatric speciation, involves cessation of gene flow between two newly formed populations as a result of geographical isolation (i.e. mountain ranges, rivers). The second model, parapatric or sympatric speciation, involves cessation of gene flow between two populations with overlapping geographical range (Bolnick and Fitzpatrick 2007; Fitzpatrick et al. 2008; Niemiller et al. 2008). Many recent genetic studies, on organisms as diverse as fish (Terai et al. 2006), birds (Ellegren et al. 2012), insects (Hearn et al. 2013; Martin et al. 2013), amphibians (Niemiller et al. 2008), mammals (Lohse and Frantz 2014; Green et al. 2010; Reich et al. 2010) and plants (Mitsui and Setoguchi 2012), have shown that genetic exchange during population diversification is more common than what was originally anticipated. Hence, the reduction of gene flow between sub-populations or species, that inhabit the same geographic range, often involves a period of extrinsic reproductive isolation before acquiring an eventual intrinsic reproductive isolation.

The mechanisms by which gene flow reduces between diverging populations that overlap in their range are still not very well understood. A major goal of geneticist and evolutionary biologist is to identify the mechanisms or genes and/or regions in the genome that are involved in the reduction of gene flow and eventually emergence of reproductive isolation between diverging populations. In animals, only a few genes have so far been identified to be involved in speciation, for example *Prdm9* in mouse (Mihola et al. 2009), and Odysseus-site homeobox (Perez and Wu 1995), *JYalpha* (Masly et al. 2006) and *GA19777 Overdrive* (Phadnis and Orr 2009) in Drosophila. These sparse examples of identified speciation genes do not

seem to suggest a common or general universal pathway/process leading to speciation but rather point to the involvement of a variety of different mechanisms in the evolution of pre- and postzygotic barriers between different species.

Speciation with gene flow could be achieved through the reduction of gene flow at specific loci in the genome, also coined islands of speciation (Turner et al. 2005; Noor and Bennett 2009). Multiple studies have successfully identified possible islands of speciation in the genome of diverging species (Turner et al. 2005; Ellegren et al. 2012), however the exact contribution of these regions in speciation is still to be unraveled. Furthermore, these studies have mainly focused on genetic variation due to single nucleotide polymorphisms (SNPs) and very few studies have investigated the role that structural variations (SVs) play in the process of population diversification (Michel et al. 2010; Vicoso and Bachtrog 2013). Copy number variations (CNVs), a class of SVs, can be a major mechanism driving gene and genome evolution by duplicating and deleting segments of the genome and as a result, create novel gene functions, disrupt gene functions, or affect regulatory mechanisms in the genome. The majority of inter-species CNV studies have focused on primates (Newman et al. 2005; Popesco et al. 2006; Dumas et al. 2007; Perry et al. 2008; Dennis et al. 2012) and suggested that species-specific copy number (CN) can be evolutionarily favored because of their adaptive benefits (Popesco et al. 2006; Dumas et al. 2007; Perry et al. 2007; Nguyen et al. 2008; Guerrier et al. 2009; Dennis et al. 2012). However, these studies only provide insights into the role of CNV between already well-diverged species (i.e. Chimpanzees and Humans), making it difficult to determine whether these variations between species have arisen during speciation or rather accumulated post-speciation.

The species of the genus Sus provide a good model to study the effect of CNV regions (CNVRs) in the process of speciation. Genus Sus comprises of at least 7 morphologically and genetically well-defined species (Frantz et. al. 2013), that

inhabit the five biodiversity hotspots in Island and Mainland South East Asia (ISEA and MSEA) (Myers et al. 2000). Recent findings showed that these species diverged during the late Pliocene (4-2.5 Mya), due to their isolation on different islands of ISEA and underwent multiple rounds of small scale inter-specific hybridization during the glacial periods of the Pleistocene (2.5-0.01 Mya) (Frantz et. al. 2013). Indeed, the frequent occurrence of glacial periods during the Pleistocene, resulted in land bridges between ISEA and MSEA allowing migration between islands (Frantz et. al. 2013). Therefore, the process of divergence between the pigs in ISEA and MSEA, effectively follows alternating periods of allopatric (warm periods) and parapatric (glacial periods) conditions. However, while these species can be identified based on morphology and/or DNA and are still capable of producing fertile offspring (Blouch and Groves 1990), the mechanisms that prevented these species from large scale homogenizing during the numerous glacial periods of the Pleistocene remain unclear.

In this study, we analyzed the complete genome sequence of 4 different species of the genus *Sus*, that are solely found in ISEA (Sus-ISEA): *Sus barbatus* (Bearded pig on Borneo), *Sus celebensis* (Sulawesi warty pig), *Sus cebifrons* (Philippine warty pig), *Sus verrucosus* (Javan warty pig) and 3 populations of the species *Sus scrofa* from Europe, China and Sumatra. We compared and contrasted the pattern of CNVs among population/species, in order to investigate the role that CNVRs may play in this on-going process of speciation.

3.2 Results

Whole genome re-sequencing data were obtained for seven populations (two individuals of the same species from ISEA; *Sus cebifrons, Sus celebensis, Sus verrucosus* and *Sus barbatus* (in case of *Sus barbatus* we obtained data from four individuals) and two individuals each from three diverged populations of *Sus scrofa*; from Sumatra, China and Europe (Table 3.1, Fig 3.1, Supplementary Table 3.1).

Previous analyses have shown the read depth (RD) method to be an accurate method for computational detection of CN of regions throughout the genome, especially with high coverage data (Sudmant et al. 2010; Bickhart et al. 2012; Esteve-Codina et al. 2013; Paudel et al. 2013). Since our main goal was the identification of inter-population CNVRs, the two samples from the same population were combined to achieve higher RD. The combined data was used to identify inter-population CNVRs between the seven populations by aligning short reads to the *Sus scrofa* reference genome (Groenen et al. 2012, see material and methods for details). In the case of *Sus barbatus*, all possible pairwise combinations of the four individuals displayed a high level of congruence in CN detection in both intra- and inter-population comparison. To avoid bias due to sampling size and total coverage we selected two of four *Sus barbatus* individuals in order to give a read coverage comparable with the other populations studied (Supplementary Table 3.1). We tested the assumption that combining individuals from the same population would not create any significant bias due to the expected higher inter- than intra-population variation by comparing CN among and between the seven populations. We found that the copy number differences (CNDs) between pairs of individuals from different populations were significantly higher than between individuals from the same population (p-value <0.001, Wilcoxon test, Supplementary Figure 3.1A and 3.1B). Thus, combining two individuals of the same species will likely result in a higher sensitivity in calling CN with a relative minimal bias in the inter-population comparison. For each population, multi copy regions (MCRs) were defined by applying a threshold of a minimum of 6 consecutive 1 kilobase (Kb) bins that have an average CN higher than 2.5. All the MCRs were then retrieved from all populations and we then chained MCRs that were (partially) overlapping between two or more populations. We computed the CN for all chained MCRs in each population and for each MCR, the standard deviation (s.d.) of CN between the seven populations was estimated. All MCRs with a s.d. ≥0.7 were regarded as CNVRs. We identified 1408 regions, encompassing 17.83 megabases

(Mb) on the *Sus scrofa* reference genome, as CNVRs (Supplementary Table 3.2, Supplementary Figure 3.1) (see material and methods for details on detection of CN, MCR, and CNVR).

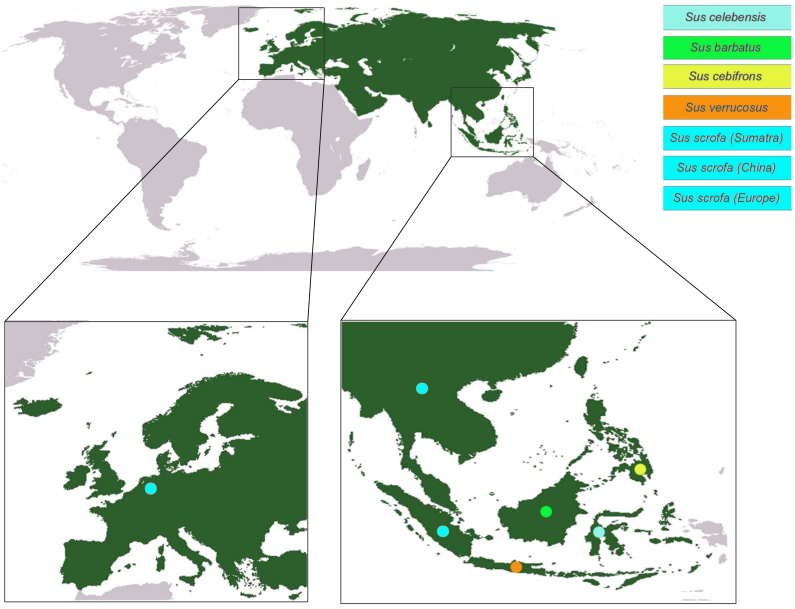


Figure 3.1 Schematic overview of origin of *Sus* populations across Eurasia and Island of South East Asia used in this study.

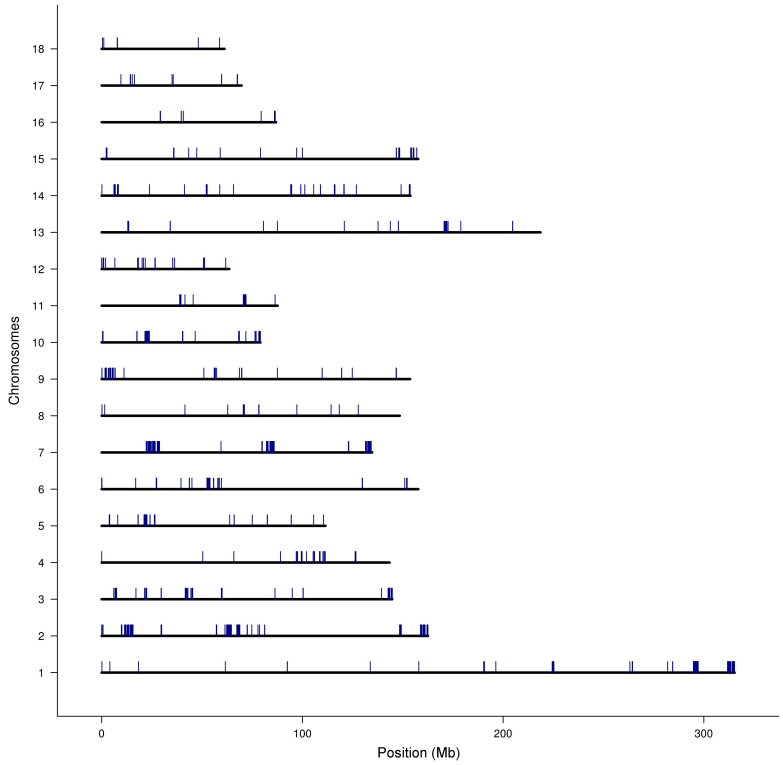
Although CNVRs were found on every chromosome, the number and the total size of CNVRs per chromosome are not correlated with chromosome length (Fig 3.2A and Fig 3.2B), which is consistent with our previous study related to CNVRs in the porcine genome (Paudel et al., 2013). Many of the identified CNVRs are relatively small, close to the effective resolution of 6 Kb. While the size of CNVRs ranges from 6 to 98 Kb, the majority (1089 out of 1408; 78%) of the CNVRs that were identified is between 6 and 15 Kb (Fig 3.2C). We did not observe any CNVR larger than 98 Kb which is probably due to incompleteness and assembly errors in the current genome build resulting in gaps in the genome. In addition, the presence of repetitive elements may preclude the chaining of smaller segments of large CNVRs. Repetitive sequences will break the contiguity of defined CNVRs as those regions

were masked in the genome prior to the alignment. We observed a number of regions on some chromosomes having cluster of CNVRs with comparatively higher CN in some populations. For example, the 0.81 Mb region between 22.24 Mb - 23.05 Mb on chromosome 10 (Fig 3.3A and 3.3B) shows higher CNs in the *Sus scrofa* populations (CN range in *Sus scrofa* 0 to 85; CN range in Sus-ISEA 0 to 39). Another example is the 370 Kb region between 78.7 Mb and 79.07 Mb on chromosome 10 (Fig 3.3A and 3.3C) that shows a series of regions with high CN in Sus-ISEA (CN range in Sus-ISEA 22 to 72; CN range in *Sus scrofa* 12 to 46).

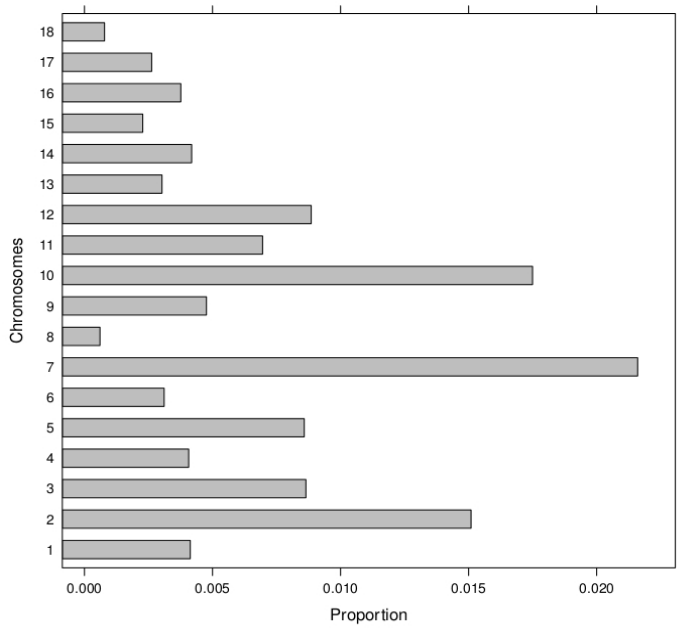
Table 3.1 Read depth of individuals and grouped individuals used (information of other *Sus barbatus* individuals can be found in Supplementary Table 3.1)

Names	Combined	Separate	Separate Depth	Combined Depth
Sus barbatus	Sbar	Sbar1	9.087	17.186
		Sbar2	8.087	
Sus cebifrons	Sceb	Sceb1	9.36	18.6
		Sceb2	9.174	
Sus celebensis	Scel	Scel1	18.409	25.475
		Scel2	7.046	
Sus verrucosus	Sver	Sver1	9.088	18.844
		Sver2	10.127	
Sus scrofa	Sumatra	Sumatra1	10.961	22.247
		Sumatra2	11.113	
Sus scrofa	China	China1	7.965	19.172
		China2	11.268	
Sus scrofa	Europe	Europe1	7.555	18.529
		Europe2	11.056	

A)


Chromosomes
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
100
200
300
Position (Mb)

B)


Chromosomes
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.000
0.005
0.010
0.015
0.020
Proportion

C)

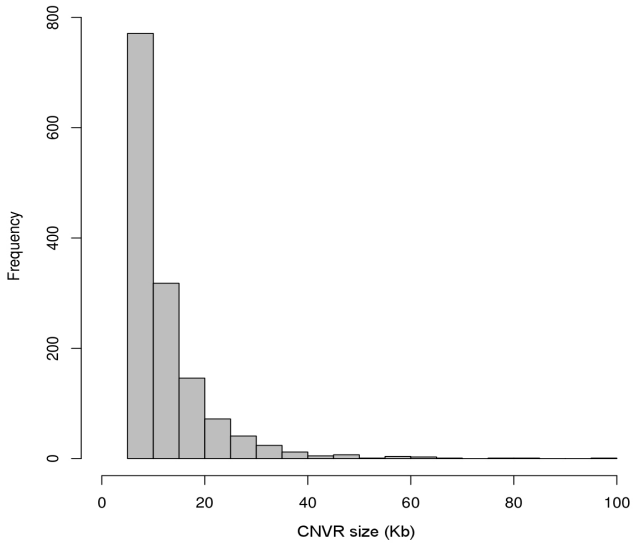


Figure 3.2 Distribution, proportion, and frequency of CNVRs in the pig genome. A) Distribution of CNVRs on the different chromosomes of the porcine genome. B) Proportion of CNVRs per chromosome. C) Frequency and size of CNVRs

Overall, most of the CNVRs identified displayed CN higher than two in all seven populations (1077 out of 1408 region) with only a small fraction (29; 211 Kb) being population specific. This could be due to the stringent criteria implemented to reduce false positive CNV calls. Sus barbatus showed the largest number of MCRs observed as variable in CN in all the seven populations (1358; 17.33 Mb) whereas Sus scrofa from Sumatra showed the lowest number of MCRs observed as variable in CN in all the seven populations (1197; 15.613 Mb) (Supplementary Table 3.3).

3.2.1 Experimental validations

We used quantitative real time-polymerase chain reaction (qPCR) to validate the identified CNVRs. We randomly selected ten genic CNVRs, ten non-genic CNVRs and five diploid regions and tested these using two distinct primer sets per locus. All 25 assays were successful and all 25 showed 100% agreement with our CNVRs predictions, indicating a low false discovery rate for calling CNVRs based on the RD analysis.

3.2.2 Functional relevance of CNVRs in the genus *Sus*

We used the porcine gene annotation of the current genome build (Sus scrofa build10.2, Ensembl release 75 (Flicek et al. 2012)) to identify genes encompassing CNVRs. To improve the reliability of the functional annotation of CNVRs, only genes having at least 70 percent overlap with a CNVR were considered. The CN of the genes were set at the CN of the overlapping CNVRs. Out of the 21,630 protein coding genes annotated in the current genome build (Groenen et al. 2012), 624 genes were found to overlap with 504 CNVRs (35.8% of total CNVRs) (Supplementary Table 3.4).

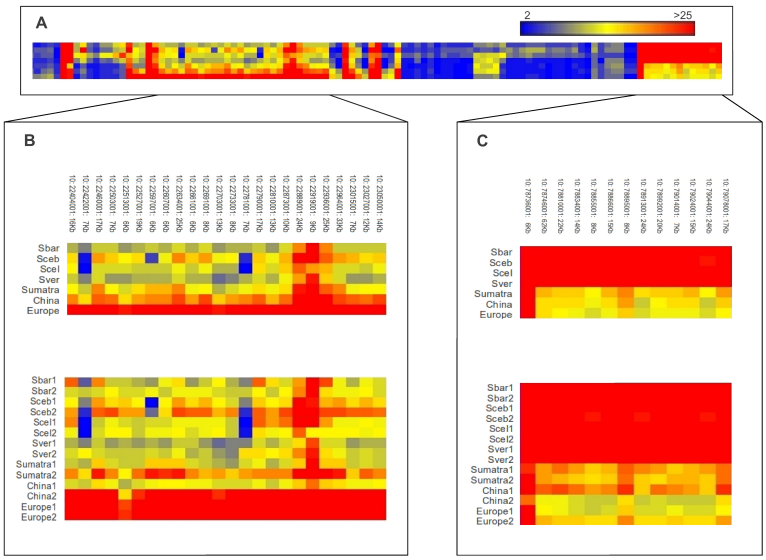


Figure 3.3 Heatmap of CNVRs. A) Heatmap of CNVRs on chromosome 10. Each column represents one CNVRs and each row represents a population. B) Heatmap of a 0.81 Mb region on chromosome 10 (SSC10: 22.24 Mb - 23.05 Mb; 24 CNVRs). Each column represents one CNVR (chromosome; CNVRs starting position; size of the CNVR) and each row represents one population (upper panel) or single individual (lower panel). C) Heatmap of a ~370 Kb region on chromosome 10 (SSC10: 78.7 Mb - 79.07 Mb; 13 CNVRs of different sizes. Each column represents one CNVRs (chromosome; CNVRs starting position; size of the CNVR) and each row represents one population (upper panel) or single individual (lower panel). Abbreviations: Sbar (*Sus barbatus*), Sceb (*Sus cebifrons*), Scel (*Sus celebensis*), Sver (*Sus verrucosus*), Sumatra (*Sus scrofa* population from Sumatra), China (*Sus scrofa* from China), Europe (*Sus scrofa* from Europe).

The olfactory receptor gene family, one of the largest gene families in the porcine genome (Groenen et al. 2012; Nguyen et al. 2012), is highly over-represented with 413 out of 624 genes overlapping a CNVR (Supplementary Table 3.4). Genes involved in immune response, such as *IFN* (Alpha-8, 11, 14; Delta-2), *IFNW1*, *IGK* (V1D-43, V2-28), *IL1B* and *PG3I*, also show variation in CN between populations.

Only few genes exhibit a high CN in a single population or a general high number of copies with much variation in two or more population. For example, *PSMB5* shows higher CNs in Sus-ISEA (from 21 in *Sus celebensis* to 10 in *Sus cebifrons*) but no sign of duplication in the three population of *Sus scrofa* (1-2 copies). *NBPF6* and *NBPF11* show high CN in all populations but with large variation in Sus-ISEA individuals (from 18 to 44 for *NBPF6* with s. d. of 11.1 and 21 to 60 for *NBPF11* with s. d. of 15.7). Likewise, *SAL1* shows CNV only between *Sus scrofa* populations (from 2-11 with s.d. of 3.48).

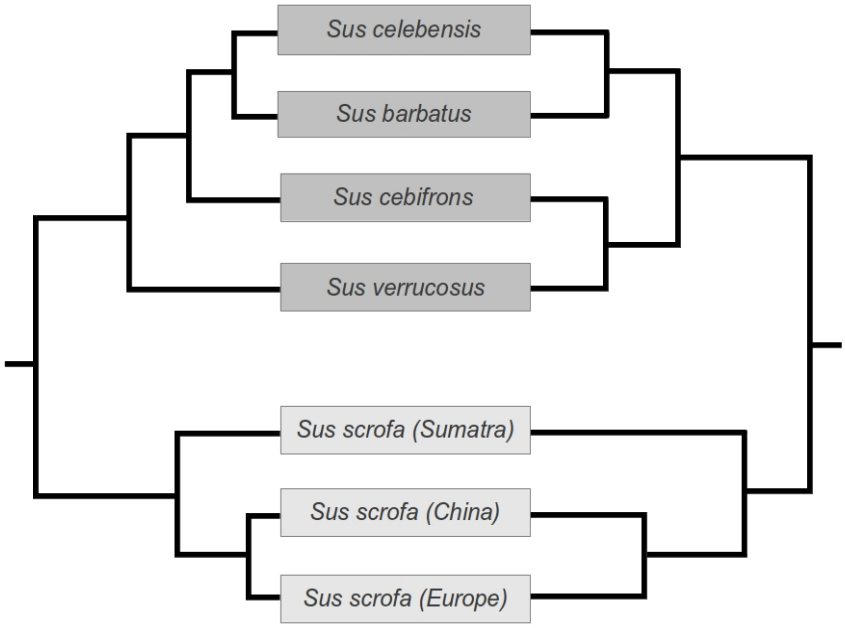
The porcine-specific immune-defense related genes *NPG3* and *PMAP23*, together with the other immune related genes *USP17L2*, *CDK20*, *POMC*, were found to be CNV with in general high variation in *Sus scrofa* populations. In addition, other previously identified CNV-genes in pigs involved in metabolism (*AMY1A*, *AMY2*, *AMY2A*, *AMY2B*) and detoxification (*UGT2B10*, *UGT1A3*, *CYPA11*, *CYPA22*, *CYP4F3,* and *CYP4X1*) are found to be CNV genes in this study as well.

A gene ontology (GO) enrichment analysis on all 624 genes overlapping CNVRs revealed that most of these genes are involved in biological processes regulating sensory perception of smell ($p<0.001$), signal transduction ($p<0.001$), neurological process ($p<0.001$) and metabolic process ($p<0.001$) (Supplementary Table 3.5).

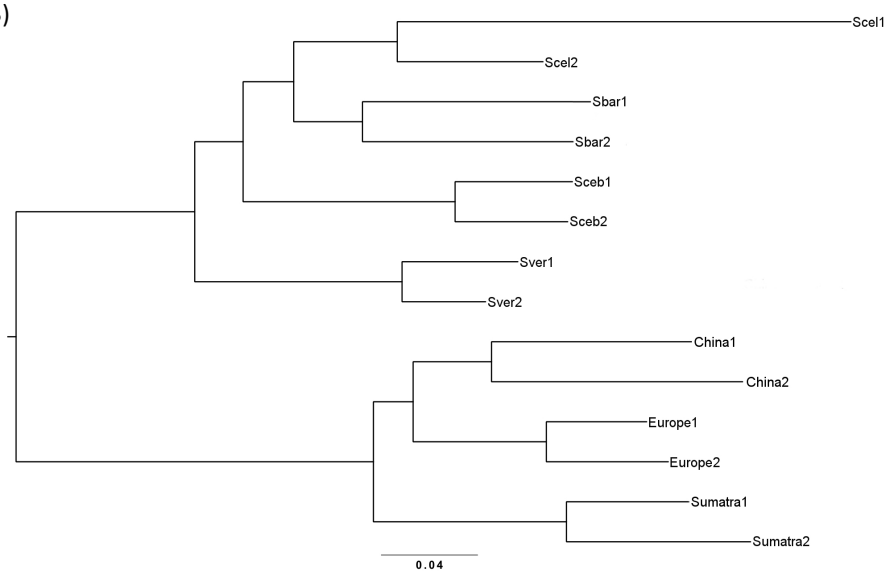
3.2.3 Cluster analysis

To investigate whether the observed CNVRs were congruent with the known phylogeny of the species, we performed a cluster analysis based on the CN at each CNV locus. The resulting tree is highly congruent to the phylogenomic analyses based on SNPs (Frantz et al. 2013) (Fig 3.4A). However, some inconsistencies are observed in the resolution of branching order within Sus-ISEA which is not surprising as recurring hybridization was common in the evolutionary history of Sus-ISEA (Frantz et al. 2013).

A)



B)



C)

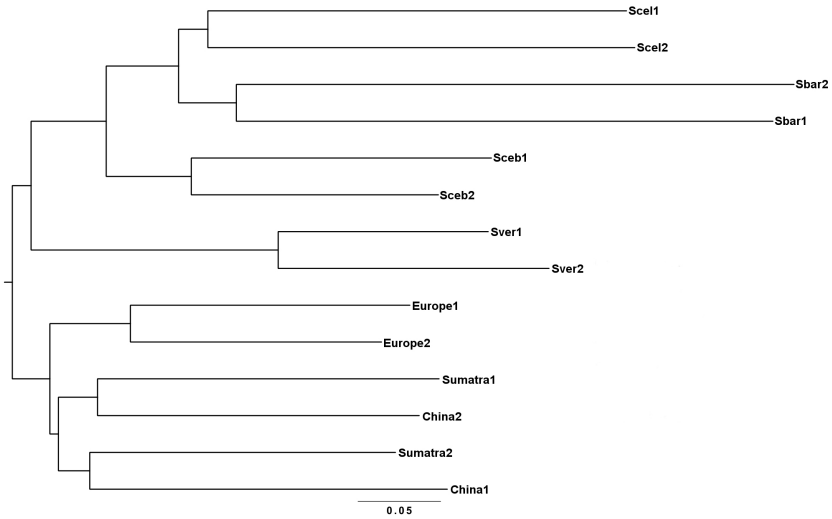


Figure 3.4: Cluster and phylogenetic tree analysis. A) Cluster analysis. The phylogenetic tree on the left side is obtained from Frantz et. al (2013) and the cluster tree on the right side is obtained by cluster analysis using the actual CN of CNVRs from different populations. The branch length does not correspond to the evolutionary distance. B) NJ-Phylogenetic tree obtained by using the pairwise difference based on SNPs (Abb. see Table 3.1). C) NJ-Phylogenetic tree obtained by using the pairwise CNDs of all possible pairs for CNVRs overlapping ORs (Abb. see Table 3.1)

3.2.4 Rate of accumulation of CNVRs (relative to rate of accumulation of SNPs)

It is generally thought that species incompatibility (e.g. through Islands of divergence) and/or lack of (intra-) species recognition are more likely to be established by fast evolving genomic regions. Thus a comparison between the rate of accumulation of CNV to other types of genetic variation, such as SNPs, could provide insight into the role of CNVs in population differentiation and speciation. To this end, a comparison between the rate of accumulation of SNPs and CNVs in each lineage was performed. To do so we first identified 1,115,908 SNPs in the genomic regions that were found to be diploid (2 copies) in all populations. We computed a rate of SNP accumulation, between each pair of individuals by dividing the number of observed difference with the total sites that could be confidentially called. Pairwise CNDs were obtained for all possible pairs of the 14 individuals. The CNDs were transformed into binary values with CND ≥ 2 as 1 and CND < 2 as 0. For each pair, the rate of pairwise difference was then calculated by dividing the total differences with the total CNVRs count (1408). The estimated CND rate is expected to be very conservative in comparison with the estimated rate of SNPs, due to our binary scale, which does not take into account the possible multiple changes in CN. For example, going from two to ten copies requires at least three duplication events but is considered as a single step in the current analysis. We observed that the rate of pairwise CND is approximately 2.5 times higher than the SNP rate (Supplementary Table 3.6 and 3.7, respectively). The observed higher CND rate compared to the SNP rate could be the result of over-representation of ORs in the list of genes overlapping with CNVRs. To investigate this, the rates of pairwise CNDs of CNVRs overlapping with ORs and without ORs were calculated separately (Supplementary Table 3.8 and Supplementary Table 3.9). In both comparisons, i.e. CNVRs overlapping with and without ORs, the rate of pairwise CNDs was observed to be higher than for SNPs. The elevated CND rate therefore does not seem to be caused solely by expansion of the OR gene family.

3.2.5 Phylogenetic analysis

The observed elevated evolutionary rate of CND may suggest that some of the CNVRs could be involved in speciation since fast evolving regions potentially play a role in the transition from pre- to postzygotic isolation. We therefore constructed neighbor joining (NJ) phylogenetic trees from SNPs and CNVRs pairwise distance matrices using PHYLIP (Felsenstein 1989). We repeated the analysis using CNVRs overlapping with OR (CNVR-OR), CNVRs overlapping with genes other than ORs (CNVR-nonOR), and all CNVRs (CNVR-ALL). Trees obtained from SNPs (Fig 3.4B) and CNVR-OR (Fig 3.4C) resulted in nearly identical topologies. The SNP-tree topology is identical to previous phylogenomic analysis (Fig 3.4A) (Frantz et al. 2013) whereas the CNVR-OR-tree topology deviates slightly form the SNP-tree in the mixed relationship of the Asian *Sus Scrofa*. By contrast, phylogenetic trees obtained from CND of CNVR-nonOR (Supplementary Figure 3.2A) and CNVR-ALL (Supplementary Figure 3.2B) resulted in different topologies compared to SNP-based phylogenies where especially the CNVR-nonOR-tree topology is highly deviating from the SNP-tree. To test if population taxon sampling plays a role in the phylogenetic results, we repeated the analysis with all pairwise combinations of the four *Sus barbatus* individuals and obtained identical phylogenetic tree topologies for all different partitions (data not shown).

To further evaluate the discrepancies between the different partitions we performed a more parametric phylogenetic approach, Bayesian phylogenetic analysis, using the MKV model (Lewis 2001) as implemented in MrBayes V2.2 (Huelsenbeck and Ronquist 2001), and an extending encoding of the CNs. We first ran the MKV model without any topology constrains and found that the monophyly of the Sus-ISEA and *Sus scrofa* clades, as identified by the SNP data and in previous analyses (Frantz et. al. 2013), was highly supported (posterior probability PP>0.9) for both CNVR-OR and CNVR-ALL, but not for CNVR-nonOR which supported a *Sus cebifrons* and *Sus scrofa* (China) relationship. To address the strength of support for

these discrepancies we tested different constrained models that fit the history of inter-specific admixture (Frantz et. al. 2013). We first computed the support (marginal likelihood; see methods) for a null model in which the monophyly of Sus-ISEA and *Sus scrofa* clades were constrained, a scenario consistent with the SNP tree. Thereafter 4 different models were tested that are described in Figure 3.5 A-D. In Model-1, we constrained *Sus verrucosus* and *Sus scrofa* Sumatra to be monophyletic (Figure 3.5A), representing known admixture among these species (Frantz et. al. 2013). In Model-2, we constrained *Sus celebensis* and *Sus scrofa* Sumatra to be monophyletic (Figure 3.5B) representing possible human translocations of *Sus celebensis* to Sumatra and neighboring islands. In Model-3, *Sus barbatus* and *Sus scrofa* Sumatra were constrained to be monophyletic (Figure 3.5C), representing known admixture between these two species/populations. In Model-4, *Sus cebifrons* and *Sus scrofa* China were constrained to be monophyletic (Figure 3.5D), representing possible migration from MSEA to the Philippines (Frantz et. al. 2013). The marginal likelihood analysis strongly supports the monophyly of the two major clade of Sus-ISEA and *Sus scrofa* for CNVR-OR and CNVR-ALL but not for CNVR-nonOR where this monophyly provides a much poorer fit. For CNVR-nonOR the difference in marginal likelihood (delta-lnL) to the null model was 7.46 (Table 3.2), which strongly supports the non-monophyly of the two major clades.

Table 3.2 Marginal likelihood scores for each partition of CNVR for different models tested.

	CNVR-ALL*	CNVR-OR*	CNVR-nonOR*
Non-constrained	7.74	7.61	6.13
Constrained (monophyly *Sus scrofa* and *Sus*-ISEA, respectively)	0	0	7.46
Constrained (*Sus scrofa* (Sumatra) and *Sus barbatus*)	47.72	16.12	21.6
Constrained (*Sus scrofa* (Sumatra) and *Sus celebensis*)	45.11	20.65	11.89
Constrained (*Sus scrofa* (Sumatra) and *Sus verrucosus*)	31.18	15.52	14.72
Constrained (*Sus scrofa* (China) and *Sus cebifrons*)	32.71	19.72	0

* delta-lnL i.e. (best marginal likelihood score) – (marginal likelihood score of the model)

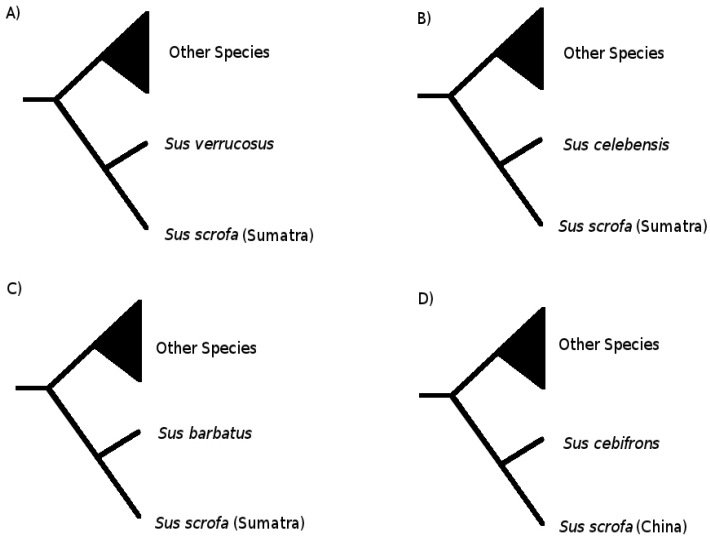


Figure 3.5 Simple schematic diagrams of tested constrained models. A) Constrained model 1 where other species consists of *Sus scrofa* (Europe and China) and *Sus barbatus, Sus cebifrons* and *Sus celebensis*. B) Constrained model 2 where other species consists of *Sus scrofa* (Europe and China) and *Sus barbatus, Sus cebifrons* and *Sus verrucosus*. C) Constrained model 3 where other species consists of *Sus scrofa* (Europe and China) and *Sus cebifrons, Sus celebensis* and *Sus verrucosus*. D) Constrained model 4 where other species consists of *Sus scrofa* (Sumatra and Europe) and *Sus barbatus, Sus celebensis* and *Sus verrucosus*.

3.2.6 *Sus scrofa* and Sus-ISEA specific CNVRs

In order to identify CNVRs specific to the two monophyletic clusters, Sus-ISEA and *Sus scrofa* (Frantz et. al. 2013), we ascertained CNVRs (s.d. ≥ 0.7) in each of these clusters separately. We found 782 and 1089 CNVRs in *Sus scrofa* and Sus-ISEA, respectively (Supplementary Table 3.10A and 3.11A). A total of 687 CNVRs were found to overlap between the two groups (ascertained as CNVRs in both group) together with 98 and 407 CNVRs uniquely ascertained in *Sus scrofa* and Sus-ISEA group, respectively (Supplementary Table 3.10B and 3.11B). We observed 243 genes in the 687 CNVRs whereas uniquely ascertained CNVRs in *Sus scrofa* and Sus-ISEA contained 47 and 178 genes, respectively (Supplementary Table 3.10C and 3.11C). Most of the genes unique to each cluster were found to be OR genes. Notable, the majority of the OR genes that were observed to vary in Sus-ISEA were found to be fixed with high CN in *Sus scrofa* populations. To test if taxon sampling introduces a bias in these group specific analyses (because of four populations in Sus-ISEA and three in *Sus scrofa*), we re-sampled every possible combination of three in the Sus-ISEA cluster. This sampling correction did not affect any of the results described above (e.g. there was always a higher number of CNVRs in Sus-ISEA than *Sus scrofa*; number of CNVRs in Sus-ISEA group varied from 917 to 1026).

3.3 Discussion

3.3.1 Evolution of CNVRs in the genus *Sus* and their possible role in the on-going *Sus* speciation process

The comparison between the seven populations of genus *Sus* allowed us to elucidate general and species-specific features of CNVs. It is known that compared to SNPs, CNVRs cover a larger part of the genome (in terms of nucleotides) and potentially have larger effects by, for example, changing gene structure, gene dosage and alternating gene regulation (Henrichsen et al. 2009; Zhang et al. 2009). In this study, we detected 1408 CNVRs in these five closely related species of the genus *Sus*. The functional enrichment analysis of the CNVRs suggested that genes

involved in sensory perception of smell, signal transduction, neurological process, and metabolic process are over-represented in CNVRs. The most abundant gene family in the porcine genome, the OR gene family, was observed as highly over-represented in the CNVRs. This over-representation of OR genes in the CNVRs could have strong functional consequences since pigs strongly rely on their sense of smell for finding food, predators, and most importantly potential mates.

The process of (on-going) speciation is thought to be triggered by a combination of many different mechanisms which include processes such as, gradual adaptation to different environment, evolution of divergent mate recognition and other molecular mechanism which are thought to be influenced by fast evolving regions in the genome. These fast evolving regions potentially accumulate divergence faster, which eventually result in creating reproductive barriers between populations. CNVRs can be a major mechanism driving gene and genome evolution by duplication and deletion of segments of the genome and as a result, create novel gene functions, disrupt gene functions, or affect regulatory mechanisms in the genome. The comparison between the rate of accumulation of CNVRs and the rate of accumulation of SNPs suggests that the CNVRs are evolving approximately 2.5 fold faster than SNPs, which is in line with a recent study in apes (Sudmant et al. 2013) where a 1.4 fold differences was observed between CNVRs and SNPs. Thus, these fast evolving CNVRs, especially those overlapping with functional regions in the genome might be a major driver of the on-going speciation in pigs.

The recent study on speciation of the genus Sus has shown that these taxa have undergone multiple rounds of small-scale inter-specific hybridization (i.e. admixture) during the glacial periods of the Pleistocene (2.5-0.01 Mya) (Frantz et al. 2013). Despite the multiple events of interspecific hybridization and being geographically very close to Sus-ISEA populations, the Sumatran *Sus scrofa* population (found to be coexisting with *Sus barbatus* on Sumatra) was found to be

less admixed with Sus-ISEA than *Sus scrofa*. This implies the existence of mechanisms that prevented these species from massive homogenizing during the numerous glacial periods of the Pleistocene. Furthermore, the phylogenetic tree analysis based on pairwise CND of CNVR-OR and pairwise difference in SNPs suggests that CNVR-OR largely recapitulates the accepted phylogeny of the genus *Sus* (Frantz et al. 2013), whereas the phylogenetic trees obtained by using pairwise CND of CNVR-nonOR, show inconsistencies with the phylogenetic history of the genus *Sus* and instead follows expected patterns of random drift and admixture (Frantz et al. 2013) (Supplementary Figure 3.2A and 3.2B). The strength of support for these inconsistencies were assessed by testing the support of different constrained models that fit the history of inter-specific admixture reported in a previous study (Frantz et. al. 2013) using a novel Bayesian phylogenetic analysis approach. The Bayesian phylogenetic analysis on the CN partitions significantly supported the recapitulations of topology of the genus *Sus* by CNVR-OR whereas for CNVR-nonOR the inconsistent topology representing admixture/random drift of genus *Sus* was strongly supported. Thus, CNVRs with OR show resistance to admixture and random drift effects between the analyzed species. This observation in combination with the observed higher rate of evolution suggests that these OR genes could play a major role in the on-going speciation process of *Sus*, facilitating rapid adaptation to different environments and divergence in mate recognition. Furthermore, pigs are known to depend highly on their sense of smell for foraging and mate recognition, and have one of the largest functional OR repertoires observed in mammals, which additionally makes it plausible that ORs are important in speciation of pigs.

Besides OR genes, genes involved in immune response, defense to pathogens and detoxification such as interferon (*IFN*), *NPG3, PMAP23* and cytochrome P450 (*CYP450*), are usually also fast evolving due to their importance for the organism to respond rapidly to changes in the environment and food-borne pathogens (Perry et

al. 2008; Liu et al. 2010; Sudmant et al. 2010; Bickhart et al. 2012; Paudel et al. 2013; Sudmant et al. 2013). Thus, together with ORs, the observed variation in CN of these genes suggests an ongoing process of evolution of these gene families and their importance for adaptation in a rapidly changing environment.

Despite the similar divergence time, the total CNVRs in the Sus-ISEA group (1089; 407 specific to Sus-ISEA) was found to be higher than that in *Sus scrofa* (782; 96 specific to *Sus scrofa*). In addition, for the 407 Sus-ISEA specific CNVRs, *Sus scrofa* shows universal high and fixed CN between three diverse *Sus scrofa* populations and most of the genes overlapping with group specific CNVRs are found to be ORs (178 genes; 146 ORs). This fixation might have happened soon after the split of the ancestral *Sus scrofa* population from the other *Sus* species from ISEA around 4 Mya.

We suggest that CNVR-ORs, might have provided the means to rapid adaption to different environments during the diversification of the genus in the Pliocene (Frantz et al. 2013). Further, the CNVR-ORs might have acted as barriers against gene flow during the multiple round of hybridization that took place later in the Pleistocene. To what extent these regions might have played a role in differentiating of *Sus scrofa* from the rest of the suids is another interesting topic which requires a more extensive taxon sampling of highly diverged suids from other parts of the world.

3.4 Materials and Methods

3.4.1 Samples and data generation

In total 16 different individuals from 5 different species were sequenced using the Illumina platform. The sequences are 100 bases pair-end reads from 400-500 bp insert-libraries with coverage per animal ranging between 7 – 18x. The sampled pigs comprised of European wild boar (2- Dutch, *Sus scrofa*), Chinese wild boar (2-

South Chinese, *Sus scrofa*), Sumatran wild boar (2- Sumatra, *Sus scrofa*), *Sus barbatus* (4 individuals), *Sus cebifrons* (2 individuals), *Sus celebensis* (2-individuals), and *Sus verrucosus* (2 individuals) (Table 3.1; Supplementary Table 3.1). Blood samples were obtained from DNA samples were obtained from veterinarians according to national legislation and tissue samples were obtained from animals culled within wildlife management programs. DNA from blood or tissue was extracted using the DNeasy blood and tissue kits (Qiagen, Venlo, NL, USA). Quality and quantity were measured with the Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

3.4.2 Sequence alignment and copy number estimation

The CN of regions in the genomes of all individuals was detected by a RD method (Alkan et al. 2009; Sudmant et al. 2010; Paudel et al. 2013), where the number of copies is inferred from sequence depth of whole genome sequence data. To calculate the average read depth from those libraries, reads were first aligned to the repeat masked reference genome (Sus scrofa build 10.2) using mrsFAST v2.3.0.2 ("Micro-read (substitutions only) fast alignment and search tool" (Hach et al. 2010)) with an edit distance of at most 7 given that the mean divergence between the seven species is maximum 2% (Groenen et al. 2012; Frantz et al. 2013). Repeat masked information was obtained from NCBI (ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Sus_scrofa/Sscrofa10.2/Primary_Assembly/assembled_chromosomes/FASTA/) and merged with the repeat masked information used in Groenen et al. (2012). Because the RD methods do not take paired-end information into consideration, all the paired-end sequences were treated as single-end sequences. Two individuals from each species were merged and treated as one to increase the confidence and sensitivity to infer CN (see results). Calculation of read depth across the whole genome was done with the help of SAMtools v0.1.18 (r982:295) (Li et al. 2009). Average read depth for each 1 Kb non-overlapping bins of repeat masked genome

was calculated. To be considered for further analysis, a bin needs to have at least 300 bases of unmasked region.

The RD method uses read depth information of diploid regions as the reference to infer CN. Since no prior information regarding diploid regions in the porcine genome was available, we initially used 1:1 orthologous genic regions between human, cow and pig and assumed these to be diploid in pig to identify CN of each 1Kb bin present in the genome. Because coding regions are known to have a higher GC content than the genome average (Högstrand and Böhme 1999; Galtier et al. 2001) this procedure may introduce a GC biased read depth. Hence, to reduce possible GC bias introduced by the 1:1 orthologous regions, all diploid regions predicted from 1:1 orthologous regions in the first stage were subsequently used to recalculate the average diploid read depth of the porcine genome as described previously (Paudel et al. 2013).

Next generation sequencing methods have been shown biased in coverage in regions of high or low GC (Bentley et al. 2008; Dohm et al. 2008; Aird et al. 2011; Benjamini and Speed 2012; Oyola et al. 2012; Quail et al. 2012). To correct for this bias we calculated GC intervals correction factors as described by Sudmant et al (Sudmant et al. 2010). These factors were then used to correct read depth of each 1 Kb bin across the genome. CN of each 1 Kb non-overlapping bin was then estimated based on the GC corrected read depth. Since the samples include both male and female individuals, sex chromosomes were excluded from the analysis.

3.4.3 Prediction of MCRs and defining CNVRs

All the 1 Kb bins with minimum CN of 1 were extracted from all individuals and bins with CN ≥2.5 were chained to form multi copy regions (MCRs). The same MCRs might be assigned with different boundaries in different individuals due to technical and/or biological reasons. Therefore, all the MCRs from all individuals

were extracted, merged, and CN of those regions for all individuals were calculated and compared. Further, the MCRs with standard deviation of CN higher than 0.7 (s.d. ≥0.7) between all individuals were assigned as CNVRs (Paudel et al. 2013).

3.4.4 Gene identification and Gene Ontology

All the annotated porcine genes from Sus scrofa build 10.2, Ensembl release 67, were extracted using Biomart (Haider et al. 2009) and genes overlapping with the CNVRs (≥70% overlap) were identified. Not all pig genes have associated gene names, thus the genes without gene names were aligned against the human Refseq mRNAs and human reference protein sequences (blastn and blastp, respectively), and the best human hit was assigned as gene name. Human orthologs of porcine genes were then used to perform a gene ontology analysis. BinGO v2.44 (Maere et al. 2005) a plugin of Cytoscape v2.8.3 (Shannon et al. 2003) was used to identify enriched GO terms using human gene annotation as background. A hypergeometric test was used to assess the significance of the enriched terms and Benjamini-Hochberg FDR correction was implemented for multiple comparisons.

3.4.5 *Sus scrofa* specific and other suids specific CNVRs

For the group comparison, we formed two groups: one with *Sus scrofa* including all three diverse populations of *Sus scrofa* and another with the Sus-ISEA. CNVRs for both groups were generated based on the similar approach described above comparing only individuals belonging to a group.

3.4.6 Cluster analysis

Hierarchical cluster analysis was performed using R package “hclust” on the CN at each CNVR. Initially, each species is assigned to its own cluster and then the algorithm proceeds iteratively, at each CNVR joining the two most similar clusters, continuing until there is just a single cluster.

3.4.7 SNP calling

We extracted all the regions that were assigned as diploid (CN 2) in all populations. We then used SAMtools mpileup (Li et al. 2009) to call genotype at sites and only considered genotype calls as SNPs, if they are different from the reference base and covered by at least 4 reads with minimum base and mapping quality of 20.

3.4.7 Estimation of pairwise distance between SNPs and CNVRs and construction of phylogenetic tree

A rate of SNP accumulation, between all possible pair of the 14 individuals was computed by dividing the number of observed difference with the total sites that could be called confidently i.e. 1,115,908 SNPs. The CNDs were transformed into binary values with CND ≥ 2 as 1 and CND < 2 as 0. For each pair, the rate of pairwise difference was then calculated by dividing the total differences with the total CNVRs count (1408). PHYLIP package (Felsenstein 1989) was used to construct neighbor joining (NJ) phylogenetic trees from the calculated pairwise distance matrix of SNPs and the following partitions of CNVRs: CNVR-OR (CNVRs overlapping OR genes) CNVR-nonOR (CNVRs overlapping non-OR genes) and CNVR-ALL (all CNVRs with and without gene overlap).

3.4.8 Construction of phylogenetic trees using a Bayesian approach

Bayesian phylogenetic analysis was performed using the MKV model (Lewis 2001) as implemented in MrBayes (Huelsenbeck and Ronquist 2001). This model implements a maximum likelihood approach to variable characters (i.e. morphology). To use this model with our CN data we need discrete CN values between 0 and 9. We used the following equation to transform CNs of each locus for each species into 9 discrete values.

CNn = ((CNo- CNmin)/(CNmax – CNmin))*(10-1)

Where, CNn = Transformed CN (rounded)

CNo = Raw CN

CNmax = Maximum observed CN for a locus

CNmin = Mainimum observed CN for a locus

We used the default (infinity) hyper-prior for the dirchelet process that model rate classes. This model implies little variation among rate of transition between CN. More complex models can be used by decreasing the hyper-prior (increasing concentration parameter). However, because increasing the concentration parameter (the number of rate categories) for the dirichelet process greatly increases the running speed, we kept this parameter to the default settings. For each data set (CNVR-OR, CNVR-nonOR and CNVR-ALL) we first ran 1,000,000 Markov Chain Monte Carlo (MCMC) (25% burnin) samples to estimate posterior distributions of the various parameters. Marginal likelihoods were computed using the stepping-stone model (Fan et al. 2011; Xie et al. 2011) with 1,000,000 samples (25% burnin) and 50 steps. We also estimated the marginal likelihood under different constrained models (see Results) to further investigate the support for discrepancies found among data sets and between NJ and Bayesian trees.

3.4.9 qPCR validation

Primer3 webtool http://frodo.wi.mit.edu/primer3/ was used to design primers for qPCR validation. Amplicon length was limited between 50 bp to 100 bp and regions with GC percentage between 30% and 60% were included, while avoiding runs of identical nucleotides. All other settings were left at their default. Details of the qPCR primers can be found in Supplementary Table 3.12. qPCR experiments were conducted using MESA Blue qPCR MasterMix Plus for SYBR Assay Low ROX from Eurogentec, this 2x reaction buffer was used in a total reaction volume of 12.5µl. All reactions were amplified on 7500 Real Time PCR system (Applied Biosystems group). The CNDs were determined by using a standard ΔCt method that compares the mean Ct value of the target CND fragments, determined from different input concentrations, compared to the mean Ct value of a known diploid reference.

3.5 Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC Grant agreement no 249894 (SelSweep project). We would like to thank the Swine Genome Consortium for the reference genome build 10.2. We thank Bert Dibbits, Animal Breeding and Genomics Centre, Wageningen University for the qPCR validation. We also thank Gus Rose and Dr. KM Schachtschneider, Animal Breeding and Genomics Centre, Wageningen University for their help during the preparation of the manuscript.

3.6 Author contribution

YP, OM, H-JM, MAMG conceived and designed the experiments. YP, OM, LAFF, H-JM performed the experiments and analyzed the data. MAMG, RPMAC contributed reagents/materials/analysis tools. YP wrote the manuscript. OM, MAMG, LAFF, H-JM, MB, RPMAC discussed and improved manuscript. All authors read and approved the final manuscript.

3.7 Additional information

Supplementary files and tables can be downloaded from this link: https://drive.google.com/file/d/0B3goLJGl6JqgTzlMOU1MNXBneDg/view?usp=sharing

References

Aird D, Ross M, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe D, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12: R18.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research 40: e72–e72.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome res 22: 778 – 790.

Blouch RA, Groves CP. 1990. Naturally occurring suid hybrid in Java. Zeitschrift für Säugetierkunde 55: 270–275.

Bolnick DI, Fitzpatrick BM. 2007. Sympatric Speciation: Models and Empirical Evidence. Annu Rev Ecol Evol Syst 38: 459–487.

Coyne JA, Orr HA. 2004. Speciation. Sinauer Associates Sunderland, MA.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell 149: 912–922.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36: e105–e105.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. Genome Research 17: 1266–1277.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silio L, Rodriguez M, Groenen M, Ramos-Onsins S, Perez-Enciso M. 2013. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics 14: 148.

Fan Y, Wu R, Chen M-H, Kuo L, Lewis PO. 2011. Choosing among Partition Models in Bayesian Phylogenetics. Molecular Biology and Evolution 28: 523–532.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

Fitzpatrick BM, Fordyce JA, Gavrilets S. 2008. What, if anything, is sympatric speciation? Journal of Evolutionary Biology 21: 1452–1459.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. Nucleic Acids Research 40: D84 –D90.

Frantz L, Schraiber J, Madsen O, Megens H-J, Bosse M, Paudel Y, Semiadi G, Meijaard E, Li N, Crooijmans R, et al. 2013. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biology 14: R107.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics 159: 907 –911.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A Draft Sequence of the Neandertal Genome. Science 328: 710–722.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Guerrier S, Coutinho-Budd J, Sassa T, Gresset A, Jordan NV, Chen K, Jin W-L, Frost A, Polleux F. 2009. The F-BAR Domain of srGAP2 Induces Membrane Protrusions Required for Neuronal Migration and Morphogenesis. Cell 138: 990–1004.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Meth 7: 576–577.

Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A. 2009. BioMart Central Portal—unified access to biological data. Nucleic Acids Research 37: W23 –W27.

Hearn J, Stone GN, Bunnefeld L, Nicholls JA, Barton NH, Lohse K. 2013. Likelihood-based inference of population history from low coverage de novo genome assemblies. Mol Ecol n/a–n/a.

Henrichsen CN, Chaignat E, Reymond A. 2009. Copy number variants, diseases and gene expression. Human Molecular Genetics 18: R1–R8.

Högstrand K, Böhme J. 1999. Gene conversion of major histocompatibility complex genes is associated with CpG-rich regions. Immunogenetics 49: 446–455.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.

Lewis PO. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. Systematic Biology 50: 913–925.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078 –2079.

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. Genome Res 20: 693 – 703.

Lohse K, Frantz LAF. 2014. Neandertal Admixture in Eurasia Confirmed by Maximum Likelihood Analysis of Three Genomes. Genetics.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21: 3448 –3449.

Mallet J. 1995. A species definition for the modern synthesis. Trends in Ecology & Evolution 10: 294–299.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research.

Masly JP, Jones CD, Noor MAF, Locke J, Orr HA. 2006. Gene Transposition as a Cause of Hybrid Sterility in Drosophila. Science 313: 1448–1450.

Mayr E. 1963. Animal species and evolution. Animal species and their evolution.

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. 2010. Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences 107: 9724–9729.

Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. 2009. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science 323: 373–375.

Mitsui Y, Setoguchi H. 2012. Demographic histories of adaptively diverged riparian and non-riparian species of Ainsliaea (Asteraceae) inferred from coalescent analyses using multiple nuclear loci. BMC Evol Biol 12: 1–15.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature 403: 853–858.

Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M, Eichler EE. 2005. A genome-wide survey of structural variation between human and chimpanzee. Genome Research 15: 1344 –1356.

Nguyen D-Q, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP. 2008. Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Research 18: 1711–1723.

Niemiller ML, Fitzpatrick BM, Miller BT. 2008. Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. Molecular Ecology 17: 2258–2275.

Noor MAF, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103: 439–444.

Oliver W. 2008. The IUCN Red List of Threatened Species.

Oyola S, Otto T, Gu Y, Maslen G, Manske M, Campino S, Turner D, MacInnis B, Kwiatkowski D, Swerdlow H, et al. 2012. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics 13: 1.

Paudel Y, Madsen O, Megens H-J, Frantz L, Bosse M, Bastiaansen J, Crooijmans R, Groenen M. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14: 449.

Perez DE, Wu CI. 1995. Further characterization of the Odysseus locus of hybrid sterility in Drosophila: one gene is not enough. Genetics 140: 201–206.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256–1260.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, et al. 2008. Copy number variation and evolution in humans and chimpanzees. Genome Res 39: 1698 – 1710.

Phadnis N, Orr HA. 2009. A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids. Science 323: 376–379.

Popesco MC, MacLaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM. 2006. Human Lineage–Specific Amplification, Selection, and Neuronal Expression of DUF1220 Domains. Science 313: 1304–1307.

Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, Bertoni A, Swerdlow H, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13: 2498 –2504.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. Genome Research 23: 1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 Genomes, et al. 2010. Diversity of human copy number variation and multicopy genes. Science 39: 641 – 646.

Terai Y, Seehausen O, Sasaki T, Takahashi K, Mizoiri S, Sugawara T, Sato T, Watanabe M, Konijnendijk N, Mrosso HDJ, et al. 2006. Divergent Selection on Opsins Drives Incipient Speciation in Lake Victoria Cichlids. PLoS Biol 4: e433.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic Islands of Speciation in Anopheles gambiae. PLoS Biol 3: e285.

Vicoso B, Bachtrog D. 2013. Reversal of an ancient sex chromosome to an autosome in Drosophila. Nature 499: 332–335.

Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H. 2011. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. Systematic Biology 60: 150–160.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics 10: 451–481.

4

Comprehensive study on copy number variation of the olfactory receptor gene family in pigs

Yogesh Paudel[1], Ole Madsen[1], Hendrik-Jan Megens[1], Laurent A. F. Frantz[1], Mirte Bosse[1], Richard P. M. A. Crooijmans[1] and Martien A. M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, De Elst 1, 6700 AH, Wageningen, The Netherlands

Submitted 2014

Abstract

Genes encoding olfactory receptors, the proteins responsible for odorant recognition, form the largest gene family in mammals and vary considerable in copy number between species. This variation in olfactory receptor repertoire is related to the level of reliance on olfaction in the context of the ecology of species, and to genetic drift resulting in random duplication and deletion of olfactory receptor genes (ORs). Pigs (*Sus scrofa*) are among the mammalian species with the highest number of functional ORs. The pig reference genome contains 1,301 ORs, of which more than 85% are functional. This high number is probably related to the dependence of pigs on their sense of smell for foraging and mate recognition. We developed a pipeline that uses next generation sequence data and read depth based method to identify copy number variable ORs in pig genomes. The pipeline outperforms approaches based on large copy number variable regions, especially when dealing with such a large and complex gene family. Even though, this pipeline is unable to detect the exact copy numbers of ORs due to cross alignment between closely related members of ORs, it can predict the copy number variation status of each gene in the OR repertoire with a high accuracy. We further investigated the significance of selection and genetic drift in the evolution of ORs in the pig by sequencing 36 wild and domesticated pigs from Asia and Europe. We observed 751 (60%) ORs having copy number variation in the pig, the majority being functional (637). Most of the copy number variable ORs are in clusters in the genome, suggesting an important role of gene clusters in promoting the variation of copy number through non-allelic homologous recombination (NAHR). Furthermore, the higher degrees of intra- and inter-population divergence of functional ORs indicate a probable role of selection on the variation of functional ORs in the pig genome. Surprisingly the distribution of the relative copy number of non-functional ORs is significantly different from a normal distribution as expected by neutral evolution of non-functional ORs. Since, both functional and non-functional ORs reside in clusters in the genome, NAHR might have facilitated the variation of both

functional and non-functional ORs. Thus, we conclude that both selection and clusters of ORs in the genome play important roles in overall copy number variation of the OR repertoire in pigs.

Key words: structural variation, copy number variation, next generation sequencing data, olfactory receptor, read depth method

4.1 Introduction

A first step in the perception of smell is the ability to detect and discriminate different odorous compounds in the environment. Sense of smell, olfaction, is very important for many animals where it contributes to discriminate between edible and noxious foods, identifying toxic substances, marking territories, and avoiding predators (Feinstein and Mombaerts 2004; Mombaerts 2004). At the molecular level, olfaction is mediated by a conserved signal transduction cascade, which is initiated by the binding of odorants to specific G-protein coupled receptors, known as olfactory receptors (Buck and Axel 1991; Beites et al. 2005). Identification of the genome wide repertoire of ORs revealed that the number of ORs varies considerably between animals (Beites et al. 2005). Although mammals typically have a large number of ORs, the number of functional ORs varies a lot between different mammalian species and seems to follow their dependency on sense of smell (Hayden et al. 2010). For example, hominid primates, including humans, have increasingly relied on vision rather than sense of smell, which during the hominid primates evolution has resulted in halving the number of functional ORs in human compared to basal primates (Rouquier et al. 2000; Gilad et al. 2003, 2004; Hayden et al. 2010; Hughes et al. 2014). Pigs on the other hand depend heavily on their olfaction for finding food, detecting predators and potential mates, which also is reflected in the large number of ORs observed in the pig genome. The OR repertoire in the current genome build of pig (*Sus scrofa* build 10.2) comprises 1301 ORs (Nguyen et al. 2012). The majorities of them are functional (1113 functional ORs and 188 non-functional ORs) and are mainly found in clusters on different chromosomes in the pig genome (Groenen et al. 2012; Nguyen et al. 2012). Similar to human OR nomenclature, by looking at sequence similarity and phylogenetic clustering with ORs from other species, pig ORs have been classified into 17 different families and 349 subfamilies (Nguyen et al. 2012). During evolution of the pig, the OR repertoire has undergone a dynamic process of duplication, deletion, and pseudogenization to meet the ecological demand of pigs

(Groenen et al. 2012). Compared to cow (880 functional and 190 non-functional ORs (Lee et al. 2013)), the pig lineage has gained some additional 230 functional ORs since their last common ancestor, illustrating the importance of olfaction for pigs and suggesting that this process of duplication, deletion, and pseudogenization of ORs could still be ongoing. Thus, studying variation of the pig OR repertoire in multiple individuals, will help to further understand the variability and evolution of this large gene family.

Structural variations (SVs) in particular copy number variations (CNVs) refer to differences in copy number of segments of DNA between different individuals of a species. Studies have shown that CNVs play an important role in the evolution of genomes in general, and gene and gene families in particular, by facilitating the gradual process of expansion and diminution (Long 2001; Otto and Yong 2002; Kondrashov and Kondrashov 2006; Conrad and Antonarakis 2007; Kim et al. 2008; Korbel et al. 2008; Innan and Kondrashov 2010; Dennis et al. 2012). Some recent genome wide studies have reported the impact of copy number variable regions (CNVRs) on the OR repertoire (Trask et al. 1998a, 1998b; Rouquier et al. 2000; Nguyen et al. 2006; Redon et al. 2006; Korbel et al. 2007; Nozawa et al. 2007; Bickhart et al. 2012; Paudel et al. 2013; Sudmant et al. 2013). However, these studies mainly focused on generating global maps of CNVRs in the genomes analyzed and were carried out at low resolution (i.e. regions equal or larger than 6Kb).

A CNVR can range from a few bases up to several mega bases (Mb) in size and affect multiple genes, like clusters of functional and non-functional genes from the same gene family, which is often the case for ORs. Thus, with the resolution currently achieved in most CNV analysis, it is often not possible to determine whether all genes within a CNVR are indeed variable in copy number. This can

potentially result in a systematic bias and mislead conclusions about the CNV of specific functional and non-functional ORs.

To avoid such complications, we developed a pipeline which uses a read depth (RD) based method to identify copy number variation of each OR locus in the OR repertoire of the pig genome. Whole genome re-sequencing (WGS) data of 36 pigs (both wild and domestic) from the Eurasian continent were used to study the dynamics and evolution of CNVs in the largest known mammalian gene super family.

4.2 Results

4.2.1 Copy number variable ORs

There are 1,301 ORs (1,113 functional and 188 non-functional) in the pig reference genome (Nguyen et al. 2012). In this study however, we only considered the 1,270 ORs that are present on the autosomes of the pig genome (1,087 functional and 183 non-functional ORs (Supplementary Table 4.1)). We aligned WGS data of 36 pigs representing 12 different domestic and wild populations from different parts of Europe and Asia (Table 4.1) against a pseudo-reference genome (see materials and methods for detail about the pseudo-reference genome). A novel pipeline was developed which uses a RD method to estimate copy number of each individual OR and to identify copy number variable ORs among the 36 sequenced individuals (Supplementary Table 4.2A, Supplementary Figure 4.1, see materials and methods for details about the detection of copy number of each OR locus).

The OR gene family is one of the most complex gene families in the pig genome. Some ORs are highly similar (~100% identical). Based on phylogenetic analysis and similarity between sequences, ORs are classified into 17 different families and 349 different sub-families (Nguyen et al. 2012). Of these 349 subfamilies, 146 have only one member whereas the rest (203 subfamilies) have 2 or more members (Nguyen et al. 2012). Thus, it is expected that some level of cross alignment of sequence

reads from closely related members of the OR gene families/subfamilies will result in overestimation of the number of copies of ORs. To minimize the overestimation of copies due to cross alignment between highly similar OR members, we tested different mismatch percentages when aligning sequences against the pseudo-reference genome. We found that allowing a maximum of 2% mismatches was most suitable to compensate for sequencing errors, distance to the reference genome, allelic variation, and to minimize cross alignment.

Table 4.1 List of individuals with their sequence coverage

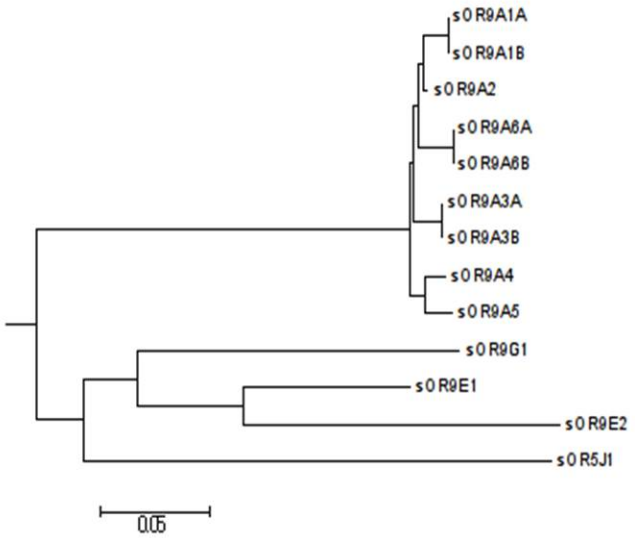
Count	Origin	Sample	Individual ID	Coverage*
1	Asian	Wild	WB20U02	7.66
2	Asian	Wild	WB29U12	8.61
3	Asian	Wild	WB29U14	8.75
4	Asian	Wild	WB29U16	12
5	Asian	Wild	WB30U08	7.51
6	Asian	Wild	WB30U09	11.11
7	European	Wild	WB21F05	7.94
8	European	Wild	WB22F02	6.63
9	European	Wild	WB25U11	9.82
10	European	Wild	WB28F31	11.59
11	European	Wild	WB42M09	11.61
12	European	Wild	WB44U07	8.29
13	Asian	Domestic	MS20U10	8.46
14	Asian	Domestic	MS20U11	8.34
15	Asian	Domestic	XI01U03	7.69
16	Asian	Domestic	XI01U04	7.51
17	Asian	Domestic	JQ01U02	7.63
18	Asian	Domestic	JQ01U08	7.24
19	Asian	Domestic	JI01U08	7.81
20	Asian	Domestic	JI01U10	8.18
21	Asian	Domestic	LSP01U16	8.86
22	Asian	Domestic	LSP01U18	10.28
23	Asian	Domestic	WS01U03	9.04
24	Asian	Domestic	WS01U13	8.4

25	European	Domestic	LW36F05	8.04
26	European	Domestic	LW36F06	8.02
27	European	Domestic	LR30F03	6.68
28	European	Domestic	LR30F04	9.06
29	European	Domestic	CM01F18	6.56
30	European	Domestic	CT01F13	9.39
31	European	Domestic	CT01M12	7.68
32	European	Domestic	MA01F20	9.29
33	European	Domestic	PI21F06	9.38
34	European	Domestic	PI21F07	10.45
35	European	Domestic	DU23M03	5.36
36	European	Domestic	DU23M04	5.91

*calculated based on diploid region in the pseudo genome

However, given the sequence similarities between family/subfamily members of the OR repertoire, this stringent criterion will not completely prevent an overestimation of copy number due to cross alignment. To test the level of cross alignment, we aligned WGS data of the reference pig (TJ Tabasco) to the pseudo-reference genome with a maximum of 2% mismatches. Without any cross alignment, we would expect the reference pig to have CN of 2 for most ORs. However, we observed that the vast majority (1127) of the ORs are estimated as having 3 copies or more in the reference pig (TJ Tabasco)(Supplementary Table 4.2B). As an example, we looked into more detail at members of the sOR9A subfamily together with its closely related subfamily members (sOR9E, sOR9G and sOR5J). These ORs have sequence distances ranging from 0.0 to 0.59 (Supplementary Table 4.2C). Figure 4.1 shows a phylogenetic tree (neighbor-joining tree) and the estimated copy number of those members in the reference individual (TJ Tabasco). We observe that members of subfamilies tend to have similar estimated copy numbers which is most likely due to the cross alignment between copies of these members. This observation further indicates that our approach cannot be used to resolve the exact copies of ORs in the genome but we presume that the bias of cross alignment will be more or less equal in all 36 individuals, enabling identification of copy number variable ORs.

We next estimated the variability of the ORs by considering an OR as variable if the standard deviation (s.d.) of the copy number of that OR in all 36 individuals was at least 0.7 (s.d. ≥0.7) (Paudel et al. 2013). Of the 1270 studied ORs, 751 (60%) were observed to be variable in copy number (CNV-ORs onwards) (Supplementary Table 4.2D) and 114 of these CNV-ORs were non-functional (62.2% of the total non-functional ORs), (Supplementary Table 4.2E) and 637 CNV-ORs were functional (58.6% of the total functional ORs) (Supplementary Table 4.2F).



ORs	Copy Number
sOR9A1A	9.16
sOR9A1B	9.16
sOR9A2	15.71
sOR9A6A	12.68
sOR9A6B	12.69
sOR9A3A	13.33
sOR9A3B	14.33
sOR9A4	10.77
sOR9A5	10.23
sOR9G1	2.62
sOR9E1	2.59
sOR9E2	2.23
sOR5J1	2.81

Figure 4.1 Neighbour-joining phylogenetic tree of 13 ORs genes (9 from sOR9A subfamily, 2 from sOR9E and 1 each from sOR9G and sOR5J subfamilies, respectively). The table shows copy number estimation of the 13 ORs on the reference individual TJ tabasco.

To test whether there is any difference in the observed copy number by considering individual OR loci (current study) instead of a larger CNVR (>= 6Kb), we compared the copy number of OR loci from this study with the copy number of ORs

from our previous study (Paudel et al. 2013). Sixteen out of the current 36 individuals were included in the previous study from which we first extracted the CNVRs overlapping with ORs from our current list. We obtained a list of 402 ORs that overlap with 297 CNVRs. Out of the 402 copy number variable ORs from the previous study, only 357 were found to be copy number variable in the current study. Similarly, 349 ORs which are assigned as CNV-OR in current study were not found as CNV in the previous study and comparing the estimated copy numbers of ORs from the two studies showed on average 200 less copies per individual in the current study (Supplementary Figure 4.2A and 4.2B, Supplementary Table 4.2G). This suggests that the whole genome analysis of copy number not only resulted in a considerable overestimation of copy number of this gene family but also incorrectly assigned some of the ORs as variable in copies.

4.2.2 Experimental validation

To validate the identified CNV-ORs, we used quantitative real time-polymerase chain reaction (qPCR). We randomly selected ten functional CNV-ORs, ten non-functional CNV-ORs and five diploid ORs and tested these using two distinct primer sets per locus. All 25 assays were successful and 23 showed 100% agreement with our CNV predictions, indicating a low false discovery rate for calling CNV-ORs based on the RD analysis.

4.2.3 The mechanism behind the variation of ORs in the pig genome

The CNV-ORs are distributed non-uniformly across the pig genome and, as expected, chromosomes in the pig genome with a large number of ORs, like chromosomes 2, 7, and 9 were found to have higher number of CNV-ORs. ORs in the pig genome are generally located in clusters (Nguyen et al. 2012). Since CNVs in different genomes are facilitated by recombination-based mechanisms (Redon et al. 2006; Sudmant et al. 2010; Bickhart et al. 2012; Paudel et al. 2013), we tested whether the variation of ORs are promoted by the non-allelic homologous

recombination (NAHR) facilitated by ORs residing in clusters. We considered two or more ORs to form a cluster if they are at most 25 Kb apart and without any non-OR gene in between. The latter was included because duplication/deletion of a non-OR gene could influence subsequent selection. Altogether, 243 clusters were observed encompassing 1,015 ORs (Supplementary Table 4.3A). Out of the 243 clusters, 187 have at least one CNV-OR. Among the 751 CNV-OR, 626 (527 functional and 99 non-functional; 83%) were found to be in those 187 clusters suggesting that the ORs in clusters are more prone to vary in copy number (p-value < 0.0001, chi-square test) (Supplementary Table 4.3B; some examples of clusters: Figure 4.2A-C).

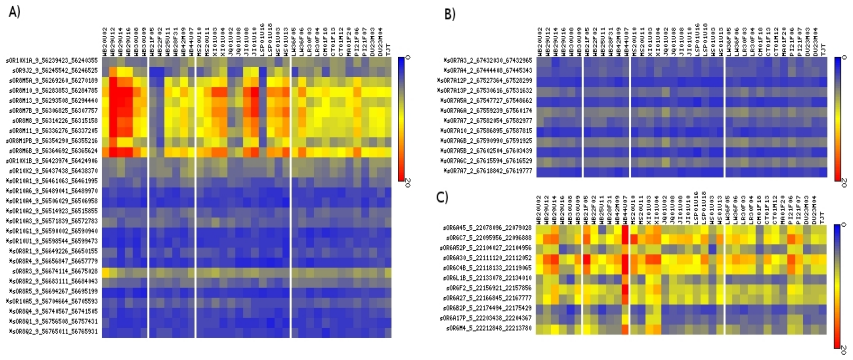


Figure 4.2 Heatmap of clusters of ORs in pig genome. A) Heatmap of cluster of ORs in chromosome 9 (SSC9: 56239423-56765931) with both CNV-ORs and non CNV-ORs (* denotes non CNV-ORs). B) Heatmap of cluster of ORs in chromosome 2 (SSC2: 67432030-67619777) where none of the ORs are variable in copy number. C) Heatmap of cluster of ORs in chromosome 5 (SSC5: 22078096-22213780) where all of the ORs are variable in copy number

Besides the cluster effect, repetitive elements such as LINEs and SINEs might also play a role in the variability of ORs. To test the role of repetitive elements in the generation of CNV-ORs in the pig genome, we examined repetitive elements in the 1 Kb flanking regions of the 1,270 OR loci. Loci that overlapped were merged to

avoid double counting and this resulted in 1,267 loci. The 1 Kb flanking sequences of these OR loci harbored a total of 5,188 repetitive elements (4.0 repeat elements/OR region)(Supplementary Table 4.3C). We then counted the number of repetitive elements in the flanking 1 Kb sequences of the CNV-ORs and nonCNV-ORs separately. The 1 Kb flanking regions of the 751 CNV-ORs harbored 2,950 repetitive elements (3.9 repetitive element/CNV-OR) while the 1 Kb flanking regions of 519 nonCNV-ORs harbored 2,240 repeats (4.3 repeat element/nonCNV-OR)(Supplementary Table 4.3D and 4.3E). These results show that there is no significant difference in repetitive element content in the flanking region of CNV-ORs and nonCNV-ORs.

4.2.4 Recently expanded ORs are more variable

Compared to cows, pigs have an additional 230 ORs suggesting OR gene expansion in the pig lineage since the last common ancestor of cow and pig. The question is if these expanded ORs are more prone to be variable in copy number or not. To test this we extracted 291 1:many and 163 1:1 orthologous ORs between cow and pig considering 1:many as “recently expanded” and 1:1 as “non-expanded" (Supplementary Table 4.4A and 4.4B). We observed 55.3% (161/291) of 1:many and 20.8% (34/163) of 1:1 ORs as variable in copy number (Supplementary Table 4.4C and 4.4D), suggesting that the recently expanded ORs are more prone to vary in copy number. Of the 161 recently expanded CNV-ORs, 151 were functional and 133 were located in clusters. In case of the 34 non-expanded CNV-ORs, 28 were functional and 30 were located in clusters.

4.2.5 Annotation of CNV-ORs

Variation in specific copies of ORs will probably alter the number of OR genes in the olfactory epithelium, ultimately altering the sensitivity to particular odorants recognized by the sensory neurons (Schaefer and Margrie 2007). Unfortunately, very little is known about the categories of odorant recognized by ORs, which

makes it difficult to elaborate on the impact of CNV-OR in a particular adaptive phenotype of pigs. ORs in pigs are classified into families and subfamilies based on their sequence identity i.e. ORs with less than 60% identity in protein sequence are classified into different families, resulting in 17 OR families (Table 4.2, (Nguyen et al. 2012)). In addition, it has been suggested that ORs with more than 60% sequence identity recognize odorants with related structures (Malnic et al. 1999; Kajiya et al. 2001), thus by comparative analysis, for some pig OR families general odor categories have been assigned (Table 4.2, (Nguyen et al. 2012)). Generally, functional ORs in the OR families assigned to be involved in mate recognition, like rancid, sour, sweat, and fatty, are less variable than functional ORs in the OR families involved in food recognition, like herbal, woody, orange, and rose (Table 4.2).

4.2.6 Comparison of variation of functional and non-functional ORs

Differences in the degree of variation of functional versus non-functional ORs might provide insight whether there is any difference in selection between the two types of ORs. To test this, average s.d. (mean and variance) of the copy number for both functional and non-functional ORs in the 36 individuals was computed. Functional and non-functional ORs were found to be significantly different in variability (Welch's t-test p-value <0.05) with an average s.d. of 1.31 and 0.9, respectively (Table 4.3). The average s.d. (mean and variance) of the copy number for both functional and non-functional ORs was computed separately for the four populations included in our data (i.e. Asian wild, Asian domestic, European wild, and European domestic). In all the analyses, functional ORs were found to be more variable than non-functional ORs and the difference in variation is significantly higher for all population except for the European wild population (Welch's t-test p-value <0.05, Table 4.3). We also combined all wild and domestic populations and calculated average s.d. (mean and variance) for functional and non-functional ORs.

Again, we observed a significantly higher variation for functional ORs compared to non-functional ORs in both wild and domestic populations (Welch's t-test p-value <0.05, Table 4.3). These observations suggest that the degrees of intra-population and inter-population polymorphism of ORs in both wild and domestic populations are higher in functional ORs compared to non-functional ORs.

Table 4.2 OR family and their function based on the sequence identity

OR Family	Members	Functional	Non-functional	CNV OR-Fun (*)	CNV OR-Non-fun	Functions
sOR11	34	23	11	6 (0.26)	7	Rancid, sour, sweaty, fatty
sOR10	64	55	9	20 (0.36)	5	Lemony, green
sOR13	62	57	5	18 (0.31)	3	NA
sOR12	12	9	3	9 (1.0)	3	NA
sOR14	15	15	0	9 (0.6)	0	NA
sOR51	43	41	2	16 (0.39)	0	Rancid, sour, sweaty, fatty
sOR53	8	8	0	5 (0.62)	0	Herbal, woody, orange, rose
sOR52	48	45	3	25 (0.55)	1	Fatty
sOR1	108	86	22	61 (0.71)	11	Vanilla, spearmint, caraway
sOR3	5	4	1	0 (0)	0	NA
sOR2	153	140	13	82 (0.58)	3	Lemon
sOR5	158	130	28	69 (0.53)	16	Fruity, spicy
sOR4	205	172	33	125 (0.72)	27	NA
sOR7	118	96	22	49 (0.51)	14	NA
sOR6	115	98	17	76 (0.78)	13	Lemon
sOR9	34	30	4	20 (0.67)	4	NA
sOR8	88	78	10	47 (0.60)	7	Floral, woody

* ratio of functional OR genes variable in copy number

Table 4.3 Divergence of functional and non-functional ORs.

	Functional ORs s.d. [*]	Non-functional ORs s.d. [*]	p-value
All	1.31 [1.38, 1.7]	0.90 [1.16, 0.80]	0.01
All wild	1.38 [1.43, 1.91]	0.94 [1.22, 0.89]	0.01
All domestic	1.30 [1.30, 1.69]	0.88 [1.08, 0.78]	0.00
Asian wild	1.28 [1.20, 1.64]	0.89 [1.04, 0.80]	0.03
Asian domestic	1.32 [1.31, 1.74]	0.96 [1.13, 0.91]	0.03
European wild	1.29 [1.31, 1.66]	0.98 [1.18, 0.97]	0.13
European domestic	1.21 [1.10, 1.46]	0.80 [0.87, 0.65]	0.00

* mean, variance

4.2.7 Does genetic drift and/or selection contribute to variation of the OR repertoire?

Various factors such as insertions, deletions, duplications, and nonsense mutations expand or diminish pseudo-genes in a genome. In general, the process of expansion or diminution of pseudo-genes (non-functional) is believed to be neutral, thus for the non-functional ORs we would expect the differences in copies to follow or approach a normal distribution whereas for the functional ORs a deviation from normality would suggest selection. To test for genetic drift/neutrality and selection in the observed CNV of ORs, we calculated relative copy number of all OR loci in each individual compared to the reference individual (TJ Tabasco, see method section, Supplementary Table 4.5A) and plotted the distributions. Figure 4.3 shows the distributions of the relative copy number for functional and non-functional ORs in all 36 individuals. We observed a similar distribution for the relative copy number of both functional and non-functional ORs. Both distributions deviate significantly from a normal distribution, however the distribution of non-functional ORs was slightly closer to normality compared to the distribution of functional ORs (Figure 4.3; Supplementary Table 4.5B and 4.5C, Supplementary figure 4.2).

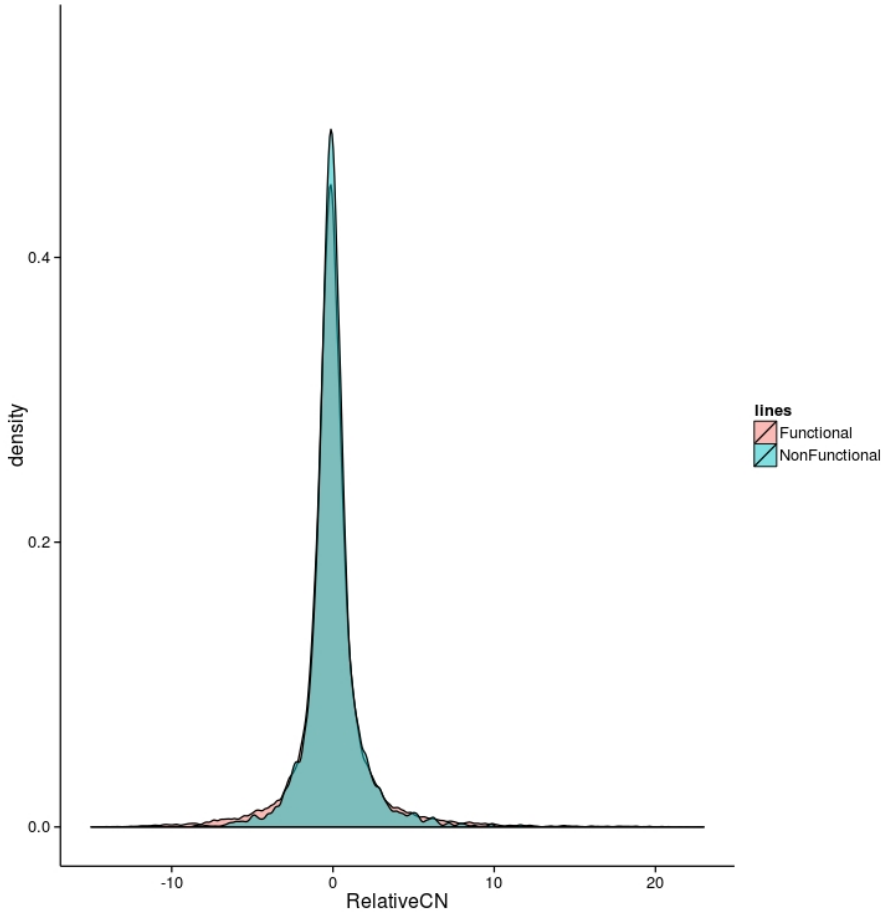


Figure 4.3 Distributions of the relative copy number for functional and non-functional ORs in all 36 individuals.

4.3 Discussion

Several genome wide studies have been performed to obtain insight in the role of CNV in the OR repertoire in a variety of mammals but most of them focused on CNVRs overlapping OR loci and not on individual ORs (Nozawa et al. 2007; Niimura and Nei 2007; Hasin et al. 2008; Young et al. 2008). In this study, we developed and used a RD approach to identify copy number variation of each individual OR gene in pig genomes. Since some members of the different OR gene families are highly

similar to each other, we tried to reduce alignments of reads from paralogous sequences by applying a stringent alignment criterion (at least 98 percent similar). However, a considerable level of cross alignment still occurs at this stringency, which is illustrated by the large number of ORs in the reference individual (TJ Tabasco) with copy number ≥ 3. This cross alignment made it impossible to resolve the exact number of copies of each OR locus using this approach. However, we assumed that the cross alignment would similarly affect all the 36 individuals. Thus, in this study we mainly focused on CNV of OR between different individual pigs. By using the pipeline we developed, we identified the status of variation of each OR locus in the OR repertoire for 36 different wild and domestic pigs from Europe and Asia, at a higher resolution than previously obtained in pigs or any other species. We observed more than half of the ORs in the OR repertoire (i.e. 751 ORs out of 1270) as having a variable copy number in the 36 pigs, with no significant differences in variation between functional and non-functional ORs (58.6% and 62.2%, respectively). The comparison of the list of CNV-ORs obtained in the current study and the CNV-ORs in the previous study (Paudel et al. 2013) suggests that around 45 out of 402 overlapping ORs were previously incorrectly assigned as being copy number variable. In addition, the previous study could not provide CNV status of the remaining 349 OR loci that were found to be copy number variable in the current study. This suggests that the pipeline presented in the current study outperforms methods based on large CNVRs (Nozawa et al. 2007; Niimura and Nei 2007; Hasin et al. 2008; Young et al. 2008) and therefore could be applied to analyze variation of ORs in other organisms and/or applied to other large gene families.

Copy number variations are known to result from a number of mechanisms such as NAHR, non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS) (Freeman et al. 2006; Bickhart et al. 2012; Sudmant et al. 2013; Paudel et al. 2013). The observed non-significant difference in frequency of repetitive

elements within 1 Kb flanking region of CNV-OR and nonCNV-OR loci, suggests that NAHR mediated by repetitive elements does not play a significant role in the variation of ORs in the pig genome. Instead, a significant number of CNV-ORs was found to reside in clusters, suggesting a prominent role of NAHR between the ORs located within clusters in facilitating the variation of ORs in pig genomes. This has previously been suggested as a mechanism in the formation of copy number variation of ORs in the human genome as well (Hasin et al. 2008; Young et al. 2008).

The pig OR repertoire has expanded by at least 230 ORs compare to its last common ancestor with cow (which has 1071 ORs, 880 functional and 190 non-functional) (Lee et al. 2013). The comparison between recently expanded and non-expanded ORs between cow and pig (see results for details on definition of expanded and non-expanded ORs) suggests that the recently expanded ORs are more prone to vary in copy number. This could be due to the lower evolutionary constraint on the newly copied genes compared to the old ORs that appear to have a fixed copy number in all 36 pigs. Although, the majority of both the recently expanded and non-expanded CNV-ORs reside in clusters, the high sequence similarity between the members of the recently expanded ORs would have favored NAHR between these ORs and thus might have promoted the observed higher copy number variation in the recently expanded ORs compared to the non-expanded ORs.

Pigs depend heavily on their olfaction for finding food and to detect predators and potential mates, which is reflected in the large number of functional ORs observed in the pig genome (Groenen et al. 2012; Nguyen et al. 2012). Almost 60% of the ORs in the OR repertoire of pigs are found to be variable in copy number, which is higher than in other organisms (Nozawa et al. 2007; Hasin et al. 2008; Young et al. 2008). In human for example, several studies have found only around 30-50% of

ORs to be variable in copy number (Nozawa et al. 2007; Hasin et al. 2008; Young et al. 2008). Functional annotation of CNV-ORs suggests that OR families involved in food detection are more variable than the ORs responsible for mate recognition (Table 4.2). This is expected because of pig's adaptation to many different environments across the Eurasian continent, which requires variation of the ORs responsible to food foraging. However, our current knowledge on OR odor specificity is still inadequate and further investigations are needed before we can draw reliable conclusion about this.

It has been suggested that positive selection could favor CNV of ORs (Nguyen et al. 2006). If positive selection favors a higher number of copies of functional ORs, then changes in the OR repertoire enhance olfactory capabilities, which gives a higher level of sensitivity to different odorants (Nguyen et al. 2006). Thus, those functional ORs are selected for as pigs adapt to a new environment. If this is the case, then the degrees of intra- and inter-population divergence of copy number of ORs should be higher for functional compared to the non-functional ORs. Supporting this hypothesis, we observed that the degrees of intra- and inter-population divergence of functional ORs are always higher (most of the cases significantly higher (Table 4.3)) indicating a role of positive selection on the variation of functional OR repertoire in the pig genomes.

In general, the process of variation of non-functional genes is believed to be mostly neutral, thus for the non-functional ORs we would expect the variation to follow or approach a normal distribution (Feller, William 1957; Nozawa et al. 2007). However, the distributions of relative copy number of non-functional ORs was not as expected (Feller, William 1957; Nozawa et al. 2007). We observed very similar distributions for both functional and non-functional ORs and both distributions deviated significantly from a normal distribution. Although, the distribution of non-functional ORs was slightly closer to normal compared to functional ORs the strong

deviation from a normal distribution suggested that other factors, in particular clustering of ORs in the genome, increase the probability of changing the number of copies of non-functional ORs facilitated by NAHR with other surrounding functional/non-functional ORs. Thus, we conclude that both selection and cluster are playing role in overall copy number variation of OR repertoire in pigs.

4.4 Materials and methods

4.4.1 Samples and data generation

In total 36 different individuals from 10 different breeds as well as wild boars from China and Europe were sequenced using the Illumina Hiseq platform. The libraries are 100 bases pair-end reads with coverage per animal ranging between 7 – 11x. The sampled pigs comprised of European wild boars (6), Chinese wild boars (6), Asian domestics (12), and European domestics (12) (Table 4.1; Supplementary Table 4.1). DNA samples were obtained from blood samples collected by veterinarians according to national legislation or from tissue samples from animals obtained from animals culled within wildlife management programs.

4.4.2 Sequence alignment and copy number estimation

Copy numbers of 1270 autosomal ORs in each individual were detected by the RD method (Alkan et al. 2009; Sudmant et al. 2010; Bickhart et al. 2012; Paudel et al. 2013), where the number of copies of each OR present is inferred from the average sequence depth of diploid region in the pig genome. For this analysis, a separate pseudo reference genome was created using 1270 ORs with 500 flanking bases. Since we need average sequence depth of diploid region in the pig genome to infer copy number of each OR locus, 700 one-to-one orthologous region between pig, cow, and human with 100 flanking bases were selected and included in the pseudo reference genome (For the reference sequence please see Supplementary File 4.1). The 500 flanking bases were used to include reads which are at the boundary of OR loci. These flanking regions were excluded for further downstream analysis. The

one-to-one orthologous regions were selected in such a way that they have similar GC distribution as that of the ORs. Because the RD methods do not take paired-end information into consideration, all the paired-end sequence libraries were treated as single-end libraries. All the sequence libraries from each individual were aligned against the reference genome using mrsFAST v2.3.0.2 ("Micro-read (substitutions only) fast alignment and search tool" (Hach et al. 2010)) with an edit distance of at most 2. mrsFAST reports all possible mapping locations for a read in the genome. We selected the edit distance 2 to minimize cross mapping between any two or more paralogous sequences in the genome.

The RD method uses read depth information of diploid region to infer copy number of each OR in the reference genome. Hence, average read depth of the 700 one-to-one orthologous regions (excluding the depth of 500 flanking bases) was used to identify CN of each OR present in the reference genome.

Next generation sequencing methods have been shown to decrease coverage of regions of high or low GC (Aird et al. 2011; Benjamini and Speed 2012; Dohm et al. 2008; Oyola et al. 2012), which is also true for other sequencing technologies (Quail et al. 2012). Polymerase chain reaction (PCR) amplification during library preparation and/or during PCR for cluster amplification on the Illumina flow-cell are known sources of under coverage of high or low G/C regions (Oyola et al. 2012). To correct for this bias we calculated G/C intervals correction factors as described by Sudmant et al (Sudmant et al. 2010). These correction factors were then used to correct read depth of each ORs in the pig OR repertoire.

4.4.3 Prediction of copy number variable olfactory receptors

Copy number variable olfactory receptors were identified based on the standard deviation of the CN of all ORs in the 36 individuals. CNV-OR status was assigned to

only those ORs, which were variable (s.d. ≥0.7) in CN across all 36 individuals (Paudel et al. 2013).

4.4.4 Alignments and Phylogenetic Analysis

Sequence alignment between the ORs was performed using default parameters in MEGA5 (Tamura et al. 2011). The number of base substitutions per site between sequences was calculated using the Maximum Composite Likelihood model (Tamura et al. 2004). The analysis involved 13 nucleotide sequences (9 members of sOR9A sub family and 4 other sequences from sOR9E, sOR9E and sOR5J sub-families). All positions containing gaps and missing data were eliminated. There were a total of 857 positions in the final dataset. Phylogenetic tree (NJ-tree) was constructed using default parameters in MEGA5.

4.4.5 Relative copy number of ORs

Copy number for all ORs in the reference individual (TJ Tabasco) was calculated as described above by using sequences obtained from the reference individual itself. The relative copy number of each OR in each individual was obtained by subtracting the copy number of OR with the copy number of the same OR in reference individual.

4.4.6 qPCR Validation

Primer3 webtool http://frodo.wi.mit.edu/primer3/ was used to design primers for qPCR validation. Amplicon length was limited between 50 bp and 100 bp and only regions with a GC percentage between 30% and 60% were included, while avoiding runs of identical nucleotides. All other settings were left at their default. Details of the qPCR primers can be found in Supplementary Table 4.6. qPCR experiments were conducted using MESA Blue qPCR MasterMix Plus for SYBR Assay Low ROX from Eurogentec. This 2x reaction buffer was used in a total reaction volume of 12.5µl. All reactions were amplified on a 7500 Real Time PCR system (Applied

Biosystems group). The copy number differences were determined by using a standard ΔCt method that compares the mean Ct value of the target CNV-OR fragments, determined from different input concentrations, compared to the mean Ct value of a known diploid reference.

4.5 Additional information

Supplementary files and tables can be downloaded from this link: https://drive.google.com/file/d/0B3goLJGl6JqgSnBmbERMVGhWRTg/view?usp=sharing

References

Aird D, Ross M, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe D, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12: R18.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.

Beites CL, Kawauchi S, Crocker CE, Calof AL. 2005. Identification and molecular regulation of neural stem cells in the olfactory epithelium. Experimental Cell Research 306: 309–316.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research 40: e72–e72.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome res 22: 778 – 790.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. Cell 65: 175–187.

Conrad B, Antonarakis SE. 2007. Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. Annu Rev Genom Human Genet 8: 17–35.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell 149: 912–922.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36: e105–e105.

Feinstein P, Mombaerts P. 2004. A Contextual Model for Axonal Sorting into Glomeruli in the Mouse Olfactory System. Cell 117: 817–831.

Feller, William. 1957. An Introduction to Probability Theory and Its Applications. Second Edition. John Wiley & Sons.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al. 2006. Copy number variation: New insights in genome diversity. Genome Research 16: 949–961.

Gilad Y, Man O, Pääbo S, Lancet D. 2003. Human specific loss of olfactory receptor genes. Proceedings of the National Academy of Sciences 100: 3324–3327.

Gilad Y, Wiebe V, Przeworski M, Lancet D, Pääbo S. 2004. Loss of Olfactory Receptor Genes Coincides with the Acquisition of Full Trichromatic Vision in Primates. PLoS Biol 2: e5.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Meth 7: 576–577.

Hasin Y, Olender T, Khen M, Gonzaga-Jauregui C, Kim PM, Urban AE, Snyder M, Gerstein MB, Lancet D, Korbel JO. 2008. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. PLoS Genet 4: e1000249.

Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Research 20: 1–9.

Hughes GM, Teeling EC, Higgins DG. 2014. Loss of Olfactory Receptor Function in Hominin Evolution. PLoS ONE 9: e84714.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97–108.

Kajiya K, Inaki K, Tanaka M, Haga T, Kataoka H, Touhara K. 2001. Molecular bases of odor discrimination: Reconstitution of olfactory receptors that recognize overlapping sets of odorants. J Neurosci 21: 6018 – 6025.

Kim PM, Lam HYK, Urban AE, Korbel JO, Affourtit J, Grubert F, Chen X, Weissman S, Snyder M, Gerstein MB. 2008. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. Genome Research 18: 1865 –1874.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. Journal of Theoretical Biology 239: 141–151.

Korbel JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M, Gerstein MB. 2008. The current excitement about copy-number variation: how it relates to gene duplications and protein families. Current Opinion in Structural Biology 18: 366–374.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318: 420 –426.

Lee K, Nguyen D, Choi M, Cha S-Y, Kim J-H, Dadi H, Seo H, Seo K, Chun T, Park C. 2013. Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. BMC Genomics 14: 596.

Long M. 2001. Evolution of novel genes. Current Opinion in Genetics & Development 11: 673–680.

Malnic B, Hirono J, Sato T, Buck LB. 1999. Combinatorial receptor codes for odors. Cell 96: 713 – 723.

Mombaerts P. 2004. Genes and ligands for odorant, vomeronasal and taste receptors. Nat Rev Neurosci 5: 263–278.

Nguyen D, Lee K, Choi H, Choi M, Le M, Song N, Kim J-H, Seo H, Oh J-W, Lee K, et al. 2012. The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. BMC Genomics 13: 584.

Nguyen D-Q, Webber C, Ponting CP. 2006. Bias of Selection on Human Copy-Number Variants. PLoS Genet 2: e20.

Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2: e708.

Nozawa M, Kawahara Y, Nei M. 2007. Genomic drift and copy number variation of sensory receptor genes in humans. Proceedings of the National Academy of Sciences 104: 20421–20426.

Otto SP, Yong P. 2002. The evolution of gene duplicates. In Advances in Genetics (ed. Jay C. Dunlap and C.-ting Wu), Vol. Volume 46 of, pp. 451–483, Academic Press.

Oyola S, Otto T, Gu Y, Maslen G, Manske M, Campino S, Turner D, MacInnis B, Kwiatkowski D, Swerdlow H, et al. 2012. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics 13: 1.

Paudel Y, Madsen O, Megens H-J, Frantz L, Bosse M, Bastiaansen J, Crooijmans R, Groenen M. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14: 449.

Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, Bertoni A, Swerdlow H, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD. 2006. Global variation in copy number in the human genome. Nature 444: 444 – 454.

Rouquier S, Blancher A, Giorgi D. 2000. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. Proc Natl Acad Sci USA 97: 2870 – 2874.

Schaefer AT, Margrie TW. 2007. Spatiotemporal representations in the olfactory system. Trends in Neurosciences 30: 92–100.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. Genome Research 23: 1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 Genomes, et al. 2010. Diversity of human copy number variation and multicopy genes. Science 39: 641 – 646.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America 101: 11030–11035.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739.

Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, Collins C, Giorgi D, Iadonato S, Johnson F, et al. 1998a. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Human Molecular Genetics 7: 13–26.

Trask BJ, Massa H, Brand-Arpon V, Chan K, Friedman C, Nguyen OT, Eichler E, van den Engh G, Rouquier S, Shizuya H, et al. 1998b. Large Multi-Chromosomal Duplications Encompass Many Members of the Olfactory Receptor Gene Family in the Human Genome. Human Molecular Genetics 7: 2007–2020.

Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. 2008. Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family. The American Journal of Human Genetics 83: 228–242.

5

Comprehensive study on short insertions and deletions in pigs

Yogesh Paudel[1], Ole Madsen[1], Hendrik-Jan Megens[1], Richard P. M. A. Crooijmans[1] and Martien A. M. Groenen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, De Elst 1, 6700 AH, Wageningen, The Netherlands

Submitted 2014

Abstract

The completion of a high quality draft genome allowed detailed investigation of a variety of population-wide genomic features in pigs. In this study, we used whole genome sequences of 42 wild and domestic pigs from Europe and Asia to generate a detailed map of short insertions and deletions (INDELs) and single nucleotide polymorphisms (SNPs). We reported over 0.5 million INDELs (size ≤10 bases) and over 6.0 million SNPs per population. The INDELs and SNPs are distributed throughout the pig genome with an average density of one INDEL per 4.6 kilobases and one SNP per 424 bases of DNA. The U-shaped distributions of INDELs and SNPs in the pig genome suggest a higher rate of INDELs and SNPs at the ends of the chromosomes. We found polymerase slippage as the major mechanism of INDEL mutagenesis in the pig genome. On average 165,000 INDELs per population were mapped to the annotated pig genes. A total of 422 coding regions were affected by INDELs resulting in a frame shift. After filtering out the wrongly annotated coding regions, 240 coding regions remained that are affected by INDELs resulting in a frame shift. Most of the coding regions harboring INDELs resulting in a frame shift were found to be annotation artifacts. Our initial observations suggested that INDELs that arise in functionally important regions of the genome, such as coding region and conserved non-coding regions, are likely to have a negative effect and are thus being removed. Hence, strong purifying selection might be driving INDEL polymorphisms between populations of pigs.

Key words: structural variation, next generation sequencing data, insertions, and deletions, INDELs, SNPs

5.1 Introduction

Structural variations (SVs) are rearrangements of chromosomes of an organism, which include insertions, deletions, duplications, inversions, and translocations. It has been shown that SVs are as important as single nucleotide polymorphisms (SNPs) in phenotypic variation and involve more base differences between individuals than SNPs (The chimpanzee sequencing and analysis consortium 2005; Mills et al. 2011; The 1000 genomes project consortium 2012).

While genome wide SNPs and large genomic SVs have been extensively studied in many different organisms, until recently very little was known about the short insertions and deletions (INDELs) and their contribution to genetic variations and influence on phenotypes (Bhangale et al. 2005). Due to improved experimental and computational strategies in the past years, it became possible to systematically catalog both short and long INDELs with high accuracy in organisms such as humans, cattle, mice, and flies (Ye et al. 2009; McKenna et al. 2010; Albers et al. 2011; Choi et al. 2013; Chong et al. 2013; Neuman et al. 2013; Montgomery et al. 2013). It has been shown that INDELs are the second largest source of pathogenic genetic variation in the human genome and that INDELs (<21 bases) account for nearly 24% of known Mendelian diseases (Mullaney et al. 2010). However, due to the lack of a high quality reference genome, comprehensive studies on INDELs have been limited to either human or model organisms such as mouse and *Drosophila* (Mills et al. 2011; Chong et al. 2013; Montgomery et al. 2013).

Insertions and deletions, especially those residing in coding regions and/or functionally important region of a genome, may cause disruption of the regulation or expression of gene, resulting in reduced fitness. Various studies have suggested that strong purifying selection acts on the removal of such variations (Mills et al. 2011; Chong et al. 2013; Montgomery et al. 2013). However, some level of positive selection might also play a role in the evolution of these variations, for instance if

an INDEL revives previously lost gene function or creates novel genes in a population. The recent completion of the pig reference genome (Groenen et al. 2012) and the advent of high-throughput-sequencing methods have allowed for a comprehensive screen of variations, including SNPs (Bosse et al. 2012; Groenen et al. 2012) and copy number variations (CNVs) (Esteve-Codina et al. 2013; Paudel et al. 2013) in the pig genomes. Although several other methods such as SNP arrays, array CGH, and read depth methods have been applied to screen for SVs in different pig populations (Kijas et al. 2001; Ramayo-Caldas et al. 2010; Ren et al. 2011; Chen et al. 2012), a comprehensive survey to detect short INDELs based on high-throughput sequences in combination with paired-end methods is still lacking. In this study, we focused on the discovery of INDELs (size: ≤10 bases) and SNPs in the genomes of 42 pigs from different populations from Europe and Asia. In addition, our aim was to investigate the evolution of INDELs in comparison to SNPs and the role of different selection processes in the evolution of INDELs in pigs.

5.2 Results

5.2.1 SNPs and INDELs discovery using NGS data from 42 pigs

To have a better understanding of SNP and INDEL evolution in pig, we examined whole genome re-sequencing data of 42 pigs from four different populations of *Sus scrofa*, covering a broad representation of pig diversity of both wild and domestic pigs from Asia and Europe (Supplementary Table 5.1). The reads were aligned against the pig reference genome (*Sus scrofa* build 10.2 (Groenen et al. 2012)) using BWA (Li and Durbin 2009). GATK (McKenna et al. 2010) was used to call INDELs (≤10 bp) and SNPs in each population (see materials and methods for more details). We applied stringent filtering steps to reduce falsely called INDELs and SNPs (see materials and methods for more details). The filtering steps included:

1) Removal of abnormally aligned paired-end reads from the alignment files.

2) Removal of falsely called INDELs and SNPs caused by assembly errors of the reference genome.

3) INDELs and SNPs within a distance of 5 bases of another INDELs and SNPs were discarded.

4) INDELs and SNPs were removed if they were fixed homozygous-alternate to the reference or heterozygous in all 42 pigs.

5) INDELs and SNPs overlapping with copy number variable regions in pig genomes (Paudel et al. 2013) were discarded.

The average sequence coverage per sample after alignment and filtering ranged from 7x to 25x (Supplementary Table 5.1). The final lists of filtered INDELs and SNPs consist of, on average, 0.58 million INDELs and 6.63 million SNPs per population (Table 5.1).

Table 5.1 INDELs and SNPs count per population.

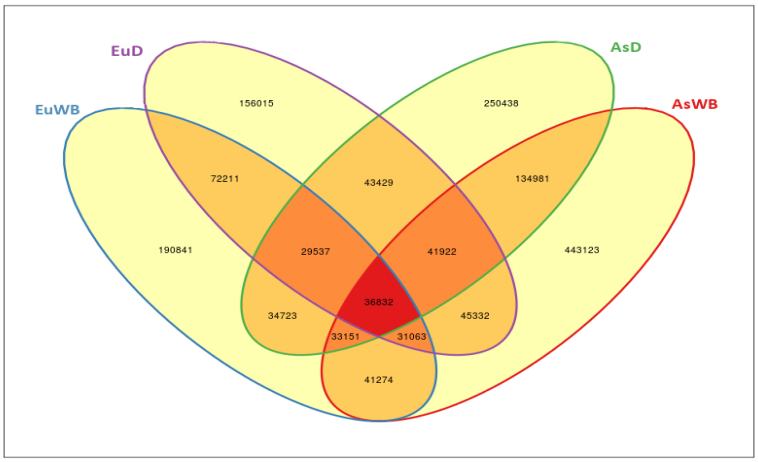
Populations	SNPs	INDELs	Insertions	Deletions
AsWB	9,669,238	807,678	346,310	461,368
AsD	7,033,394	605,013	268,289	336,724
EuWB	4,835,991	469,632	202,188	267,444
EuD	4,984,840	456,341	195,375	260,966

AsWB: Asian wild population; AsD: Asian domestic population (Meishan population from Asia); EuWB: European wild population; EuD: European domestic population (Large White population from Europe)

The smallest number of INDELs and SNPs was identified in European domestic populations and European wild populations respectively, whereas the largest number of INDELs and SNPs was identified in Asian wild populations, which is in agreement with the demographic pattern of the pig populations (Table 5.1, Figure 5.1A and 5.1B). The average distance between two INDELs and two SNPs was

around 4,688 and 424 bases respectively. The density of both INDELs and SNPs was found to be U-shaped, suggesting higher density of both INDELs and SNPs towards the ends of the chromosomes (Figure 5.2A and 5.2B). The majority of INDELs were 1 base long and for all INDEL size categories, the frequency of deletions is higher compared to insertions (Figure 5.3).

A)



B)

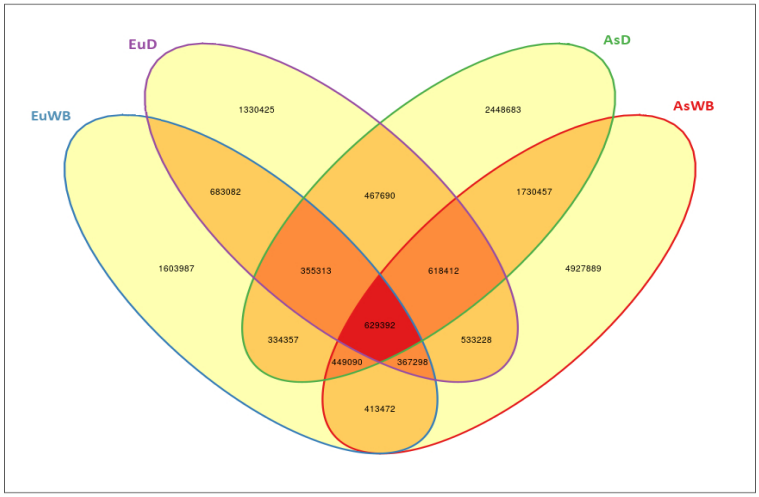


Figure 5.1 INDELs and SNPs between populations. **A)** Number of unique and shared INDELs between populations. **B)** Number of unique and shared SNPs between populations. Red eclipse represents Asian wild population, green eclipse represents Asian domestic

population, blue eclipse represents European wild population, and purple eclipse represents European domestic population

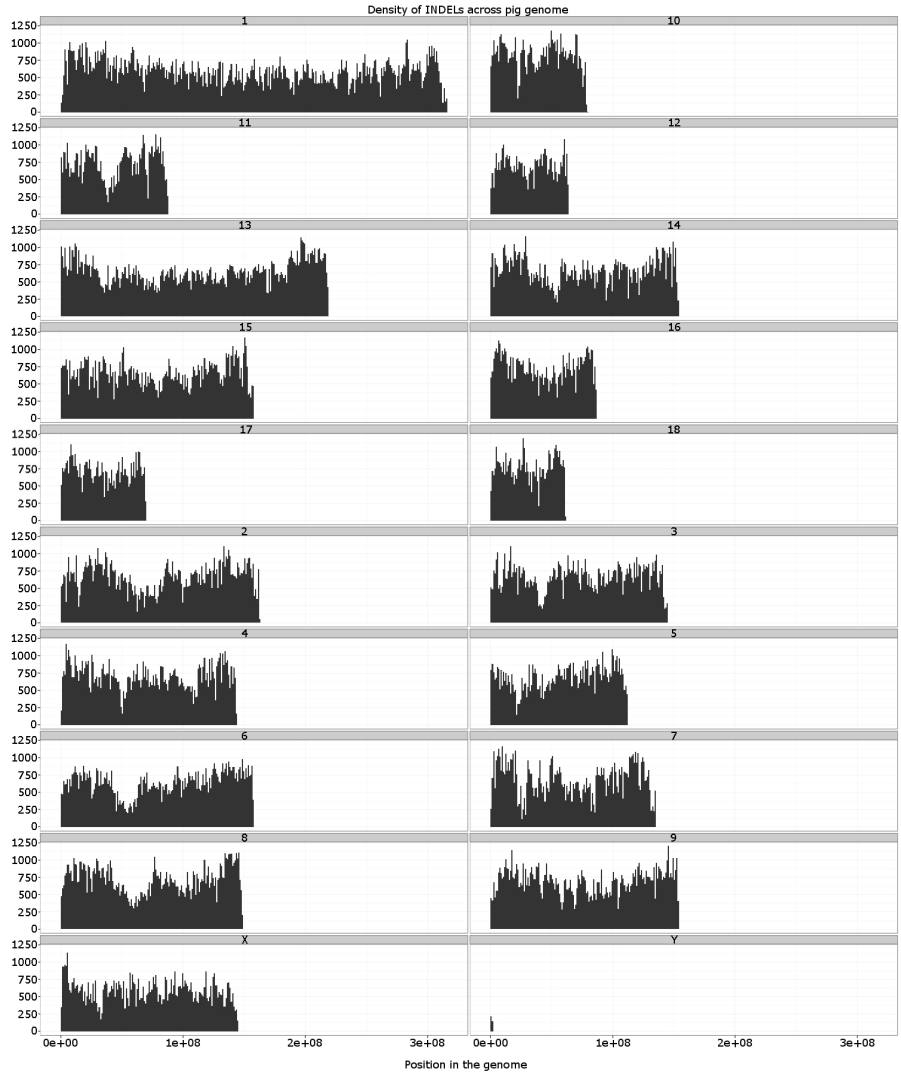
5.2.2 Association of INDELs and SNPs with different genomic features in the pig genome

To test the effect of recombination on INDELs and SNPs, for each population, we tested the correlation between recombination rate and the number of SNPs and INDELs in 1Mb segments (the size used for recombination frequency is the same as the size of recombination map currently available for pigs (Tortereau et al. 2012)). For each population, we observed a positive correlation between recombination rate with both INDELs and SNPs (Table 5.2). Similarly, for each population, in bins of 200Kb, we measured density/frequency of repeatitive elements, INDELs, SNPs, and GC percentage and found positive correlations between INDELs and SNPs with both GC content and repetitive elements (Table 5.2).

Table 5.2 Association of INDELs and SNPs with recombination rate, GC percentage, and repetitive elements frequency.

	Correlation					
	Recombination rate		GC percentage		Repetitive elements	
Population	INDELs	SNPs	INDELs	SNPs	INDELs	SNPs
AsWB	0.424	0.588	0.364	0.482	0.349	0.301
EuWB	0.269	0.472	0.182	0.327	0.212	0.201
MSAN	0.335	0.586	0.218	0.413	0.267	0.263
LWTE	0.270	0.496	0.161	0.326	0.212	0.216

A)


Density of INDELs across pig genome
1
10
11
12
13
14
15
16
17
18
2
3
4
5
6
7
8
9
X
Y
0
250
500
750
1000
1250
0e+00
1e+08
2e+08
3e+08
Position in the genome

B)

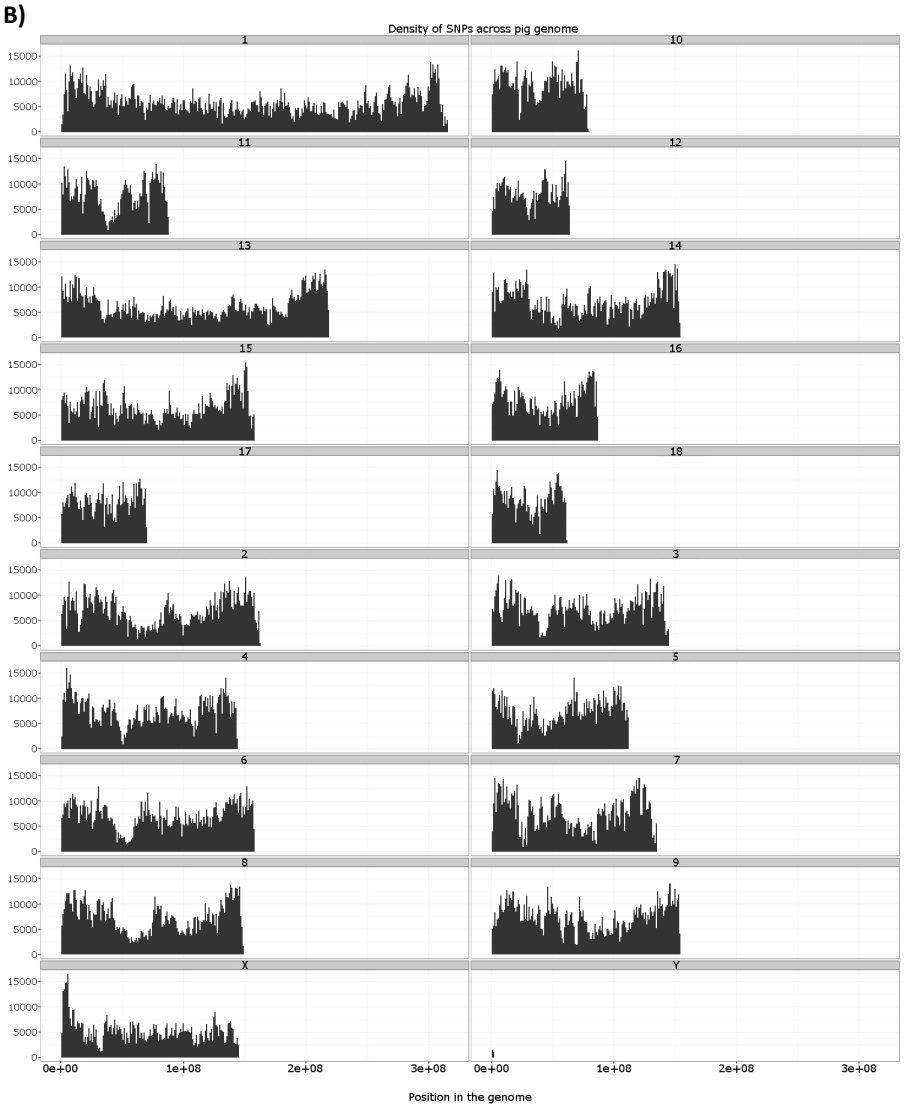


Figure 5.2 INDEL and SNP density across the pig genome (each row contains two chromosomes and density of INDELs and SNPs across the chromosomes). A) Density of INDELs across pig genome. B) Density of SNPs across pig genome. The figures represent density of INDELs and SNPs for Asian wild boar populations only. For other populations see Supplementary Figures 5.1-5.3.

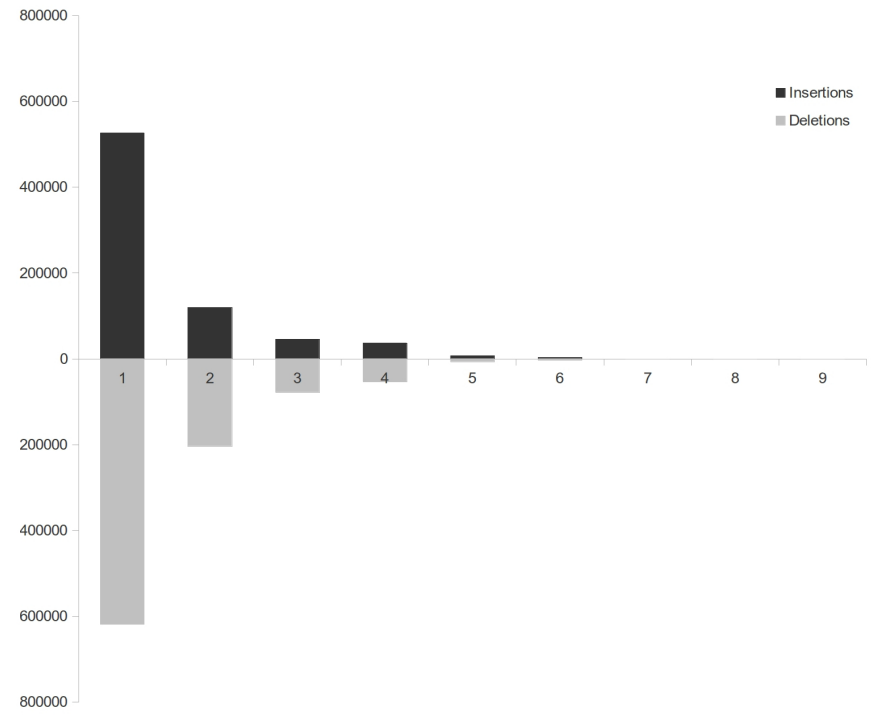


Figure 5.3 Histogram of number of insertions and deletions based on INDEL size in the pig genome.

5.2.3 Polymerase slippage and INDEL formation

To understand the role of polymerase slippage in the formation of INDELs in the pig genome, we assessed the context of the flanking regions of INDELs and categorized INDELs into five different classes as in Montgomery et al (Montgomery et al. 2013). The five classes were: 1) INDELs surrounded by homopolymer runs (HR); 2) tandem repeats (TR), if adjacent flanking regions consist of multiple repeat units similar to INDELs; 3) predicted hotspot regions (PR), INDELs adjacent/surrounded by repetitive regions and that did not fulfill the TR and HR criteria; 4) INDEL regions which did not fulfill any of the criteria above were classified into non-repetitive (NR), which was further classified into NR-change copy count (NR-CCC), if INDEL changes the local copy count of the segment, and 5) NR-nonCCC which do not

change the local copy count of the segment. In all four populations, we observed more than 80% of the INDELs as change copy count (CCC) type (HR, TR, PR, and NR-CCC) and the remaining 10-20% as nonCCC type in the pig genome reflecting the greater role of polymerase slippage in the formation of INDELs in pig genomes (Table 5.3).

Table 5.3 Types of INDELs.

	INDEL sizes and count*									
INDEL Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HR	236004	24964	3816	955	133	68	0	0	0	0
TR	249924	59233	18070	10965	816	0	0	0	0	0
PR	32708	4276	1355	745	41	14	0	0	0	0
NR-CCC	25439	34246	10392	10318	622	758	0	0	0	0
NR-nonCCC	32694	20124	13532	13323	1519	622	1	0	1	0

*Types and size of INDELs with their count in Asian wild boar population only.

5.2.4 INDELs and SNPs in annotated pig genes

To assign potential functional roles to the identified INDELs and SNPs, genes overlapping with these variants were identified. We used the porcine gene annotation of the current genome build (*Sus scrofa* build10.2, Ensembl release 75 (Flicek et al. 2012)). Around 29% of both the INDELs and SNPs in the pig genome were found to overlap with annotated genes and around 0.3% of those variations were found to reside in the coding regions (Table 5.4). Among the INDELs in coding regions, around 75% was found to cause a frame shift, which could lead to a premature termination of the gene product and potentially alter gene functions (Table 5.4, Supplementary Tables 5.2A-D). The remaining 25% of the INDELs that affect the coding regions were multiples of 3 nucleotides and thus, resulted in the precise insertion or deletion of one or more codons.

Table 5.4 Functional annotations of INDELs

Population	INDELs	Genic	CDS	3' UTRs	5' UTRs	INDELs*	Population Specific INDELs*	Genes affected by INDELs*
AsWB	807,678	233,612	638	7670	19338	490	246	233
MSAN	605,013	168,867	323	5548	14124	231	83	78
EuWB	469,632	131,395	361	4388	10780	286	68	64
LWTE	456,341	126,448	251	4208	10172	187	48	47

* Resulting in a frame shift

5.2.5 Evidence for purifying selection on INDELs in coding regions and highly conserved non-coding regions

Purifying selection removes mutations that cause reduced fitness of a population. It has been observed that the majority of all the exonic mutations are negatively selected in a population and there are very few examples where genes can tolerate structural changes without greatly affecting the function of a protein (Mullaney et al. 2010; Mills et al. 2011; Chong et al. 2013; Montgomery et al. 2013). Hence, to test the strength of selection acting on INDELs in coding region in pigs, we compared the distribution of INDELs to that of SNPs in different regions of the genome. The genome-wide ratio of INDELs to SNPs was observed to be approximately 0.09. We observed similar ratios for 3'-UTR, 5'-UTR, promoter, intronic and inter genic regions. However, the ratio for coding exons (~0.01) was well below the genome wide average (~0.09). This is most likely due to the reduction in the INDEL counts in these regions compared to the other parts of the genome.

Comparative sequence analyses between different mammalian genomes have revealed the existence of highly conserved non-coding sequences under strong purifying selection (Katzman et al. 2007; Takahashi and Saitou 2012). These regions tend to reside in gene-poor regions of the genome and are often associated with

developmental genes that may be megabases away (Bejerano et al. 2004). To test the role of purifying selection on INDELs and SNPs in the pig genome, we compared the distribution of INDELs and SNPs in constrained elements. Similar to that of the coding regions in the genome, the ratio between INDELs and SNPs in constrained elements (~0.07) was found to be below the genome wide ratio of INDELs and SNPs (~0.09) indicating a reduction of INDELs compared to the non-constrained/non-conserved part of the genome.

5.2.6 INDELs are likely to change gene function

To further understand the possible contribution of INDELs to phenotypic variation, we focused on the INDELs likely to cause disruption of gene function. For that, we selected INDELs that were specific for one of the populations and further categorized these into INDELs resulting in a frame shift and INDELs that do not result in a frame shift. The group of Asian wild boar have the largest number of INDELs in coding regions i.e. 638 INDELs. Among them 490 INDELs are group specific and 344 INDELs result in a frame shift (Table 5.4, Supplementary Tables 5.2A-D). The European domestic group was found to have the lowest number of INDELs in coding regions. This group has 251 INDELs in coding regions and among them 48 of which are specific to European domestic where 47 result in a frame shift (Table 5.4, Supplementary Tables 5.2A-D). The majority of the genes in the pig genome were annotated automatically by searching for the features of genes such as transcripts and using sequence similarity with other mammals, while some were manually curated, by reviewing the identified transcripts on a case by case basis (Flicek et al. 2012). Hence, to avoid considering falsely annotated genes as affected by INDELs that result in a frame shift, we designed a test to assess the quality of the annotation of such genes. The test includes two steps, 1) pairwise nucleotide sequence comparison of all the coding regions affected by INDELs resulting in a frame shift with their human orthologs (for human-pig comparison). Similarly, for the same set of coding regions in human, we extracted nucleotide sequence of

mouse orthologs and performed pairwise nucleotide comparisons (for human-mouse comparison) (more details in materials and methods section). 2) Pairwise nucleotide comparison between segments containing INDELs with 30 flanking nucleotide bases in the affected coding region in pigs and orthologous regions from human (for human-pig comparison) and mouse (for human-mouse comparison). We then considered only those coding regions for which the human-pig sequence identity was at least 50% and for which the ratio of the human-pig sequence identity and the human-mouse sequence identity was at least 0.8. Of the 422 coding regions affected by INDELs resulting in a frame shift, 240 fulfilled these criteria (Supplementary Tables 5.2E-H).

In the absence of specific phenotypic information for the samples used, the exact effect on the phenotype of the INDELs resulting in a frame shift cannot be firmly assessed. However, numerous examples of INDELs resulting in frame shift in genes that have phenotypic consequences have been described in pigs, humans, and mice (Kerem et al. 1989; Kijas et al. 2001; Ogura et al. 2001; Lugassy et al. 2006; Raeder et al. 2006). We, therefore, investigated which genes are harboring INDELs that result in a frame shift and that have been studied in other species. One example is the *MC1R* gene, where an insertion of two bases (CC) at position 67 resulted in a frame shift, was observed in European domestic pigs (Large White). Genes involved in growth such as *MUSK*, *ADRB3* were found to be affected by INDELs resulting in a frame shift in Asian domestic group (Meishan). Since RNA-seq data were available (unpublished results) for some of the individuals of the European domestic pig (Large White), we checked whether the INDELs are in coding region or not. Most of the INDELs resulting in a frame shift in European domestic pig (28 out of 30) were found in wrongly annotated regions of genes (Supplementary Table 5.3).

5.3 Discussion

In this study, we used high-throughput genome sequences of 42 wild and domestic pigs from Europe and Asia to generate a detailed map of small INDELs (size ≤10 bases). Studies in other mammals such as human, mouse, and cow have shown that SNPs and INDELs are abundant in mammalian genomes (Mills et al. 2006; Lunter 2007; Mullaney et al. 2010; Mills et al. 2011; Choi et al. 2013; Chong et al. 2013; Montgomery et al. 2013). However, SVs in general and INDELs in particular have received relatively little attention in pigs due to the difficulties in reliably detecting INDELs compared to SNPs. We have identified over 0.5 million INDELs (size ≤10 bases) and over 6.0 million SNPs per population. The proportion of INDELs detected in this study accounts for around 8.8% of all observed polymorphisms, including SNPs. This indicates that similar to SNPs, INDELs are not only abundant in pig genomes but also are likely to contribute to both genomic and phenotypic diversity in pigs. The observed distribution of INDELs and SNPs in the genome (average distances between two INDELs and between two SNPs were around 4,688 and 424 bases, respectively) will be affected by our method used to identify and filter variations as well as the size-range of INDELs considered. However, the overall observed distributions of INDELs and SNPs were similar to that seen in other organisms (Petrov et al. 2000; Bhangale et al. 2005; Mills et al. 2006; Mullaney et al. 2010; Mills et al. 2011; Montgomery et al. 2013).

As in other organisms, INDELs and SNPs in pigs were not equally distributed throughout the genome. The chromosomal distributions of INDELs and SNPs were U-shaped suggesting a higher rate of INDELs and SNPs towards the end of the chromosomes. A similar distribution was observed for the recombination frequency in pigs and SNP variation in previous studies (Tortereau et al. 2012; Bosse et al. 2012). We observed positive correlations between GC content, recombination frequency, and repetitive elements frequency with the INDELs suggesting some role of replication, repair, and/or recombination-based mechanisms on creating

some of these genetic variations in the pig genome. Similarly, we observed that over 80% of all INDELs in pigs are consistent with being caused by polymerase slippage across the pig genome (Table 5.3), which is consistent with the observation in human genome (Montgomery et al. 2013).

On average, more than 165,000 INDELs per population were found in the annotated pig genes, and 422 coding regions on average saw INDELs resulting in a frame shift (Table 5.4; Supplemental Tables 5.2A-D). After filtering out putative wrongly annotated coding regions, we obtained 240 coding regions affected by INDELs resulting in a frame shift (Supplementary Table 5.2E-H). This collection of INDELs and SNPs represents a valuable resource for future studies on the relation between genomic variation and phenotype. Many of the INDELs in coding regions are expected to alter gene function (Table 5.4; Supplemental Tables 5.2A-H). Several studies have shown that INDELs can confer pathogenic alterations to gene function and sometimes result in diseases or susceptibility to diseases. For instance; maturity-onset diabetes in human is caused by a single base deletion in *CEL* gene (Raeder et al. 2006). Cystic fibrosis is caused by deletion of three bases in the *CFTR* gene (Kerem et al. 1989). Ectodermal dysplasia syndrome is caused by a deletion in the *KRT14* gene (Lugassy et al. 2006). A frame shift mutation in *NOD2* gene is associated with the susceptibility to Crohn's disease (Ogura et al. 2001). Similarly, some studies have identified INDELs within coding regions that cause phenotypic alterations in pigs as well. Examples are a deletion of 6 bases in the *TYRP1* gene causing the brown coloration phenotype in a Chinese pig breed (Ren et al. 2011) and a frame shift deletion in the coding region of the *MC1R* gene, shown to be involved in the recessive red coat color phenotype in pigs (Kijas et al. 2001). However, the RNA-seq data analysis of the INDELs in European domestic population suggests that most of the coding regions affected by INDELs resulting in frame shift in fact are artifacts of the annotation. Since we have found many genes with INDELs resulting in frame shift in all populations, future studies involving gene

expression analyses should be conducted to verify these disruptive INDELs to understand their role in phenotypic variation in pigs.

Our observations suggest that INDELs that arise in functional regions appear to be eliminated by strong purifying selection. This is especially true for the coding regions in the genome. This is due to the general deleterious effect of INDELs that result in a frame shift. In addition, analyses of the INDELs in conserved non-coding regions suggest that compared to SNPs, INDELs in the conserved non-coding regions are strongly eliminated by purifying selection. This is in line with other studies, which have shown that highly conserved non-coding sequences are under strong purifying selection in mammalian genomes (Katzman et al. 2007; Takahashi and Saitou 2012).

All these observations indicate that INDELs, especially those residing in functionally important regions of the genome, are likely to have a negative effect and are thus being removed. Hence, strong purifying selection might be driving INDEL polymorphisms between populations of pigs.

5.4 Materials and Methods

In total 42 different individuals originating from 7 different populations of *Sus scrofa* were sequenced using the Illumina HiSeq platform. Sequences were 100 bases pair-end reads from 400-500 bp insert-libraries with a genome coverage per animal ranging between 7 – 25x. The sampled pigs comprised of European wild boar (5- Dutch wild boars, 8- Italian wild boars), Asian wild boar (2- North Chinese wild boars, 2- South Chinese wild boars and 8- Japanese wild boars), European domestics (9- Large Whites), and Asian domestics (8- Meishan) (Supplementary Table 5.1). DNA samples were obtained from blood samples collected by veterinarians according to national legislation or from tissue samples obtained from animals culled within wildlife management programs.

5.4.1 Sequence alignment

Since we obtained sequences from different Illumina sequencing platforms, we used SeqRet (EMBOSS tool (Rice et al. 2000)) to convert the quality scores (Phred+64) of reads to Sanger quality (Phred+33) if needed. Reads were quality trimmed using Sickle v1.2 (Joshi and Fass 2011) with default parameters where both reads from a paired-end reads were discarded if the length of any one of trimmed pair-end read was less than 30 bases. Trimmed reads were then aligned using BWA (aln) v0.6.1-r104 (Li and Durbin 2009) to the pig reference genome (Groenen et al. 2012)). BWA utilizes backward search with Burrows–Wheeler transform (BWT) and has been shown to be suitable and optimal for both SNP and INDEL calling (Li and Durbin 2009). We allowed a maximum of 7% of the bases of a read to have mismatches (-n option in BWA to 0.07). Alignments were filtered and uniquely aligned reads were kept for further analysis (using SAMtools (Li et al. 2009) and Picard v1.95; http://picard.sourceforge.net).

5.4.2 INDEL and SNP calling and primary filtering

INDELs of size less than 10 bp and SNPs were called using GATK v2.6.5-gbd531bd (McKenna et al. 2010; DePristo et al. 2011). To improve the quality of INDELs called we used the local re-alignment step IndelRealigner in GATK. UnifiedGenotyper variant caller (GATK V2.6.5-gbd531bd) was used to call SNPs and INDELs in each population. We filtered out all the SNPs and INDELs with mapping quality (MQ) less than 45 and with quality by depth (QD) less than 4 (GATK V2.6.5-gbd531bd).

5.4.3 Secondary filtering of INDEL and SNP

To avoid false calling of INDELs and SNPs, we removed all the INDELs and SNPs, which have another SNPs or INDELs or both within 5 bases flanking INDEL/SNP region. Problems in the reference genome could also cause false calling of INDELs. To avoid these false positives, we called INDELs and SNPs on the reference

individual itself (individual name: TJ Tabasco, sequence depth 25x, alignment and SNP and INDEL calling was done in the same way as described above). We extracted all the homozygous INDELs and SNPs from the reference individual and removed all the INDELs and SNPs from all the populations, which either overlap or reside within the vicinity of 5 bases of homozygous INDELs and SNPs from the reference individual. Further, we extracted fixed homozygous and heterozygous INDELs and SNPs in all 42 samples and removed them from the list of INDELs and SNPs. Finally, INDELs and SNPs that did overlapping with copy number variable regions (CNVRs) were discarded. For that the CNVR list from our previous analyses was applied (Paudel et al. 2013).

5.4.4 Filtering wrongly annotated Genes

We extracted all the orthologous coding regions (nucleotide) in human and mouse using the BioMart tool in the Ensembl genome browser (release 75, (Flicek et al. 2012)). We performed pairwise local alignment of nucleotide sequence of all the coding regions in pigs affected by INDELs resulting in a frame shift with their human orthologs (human-pig comparison). For the same set of coding regions in human, we extracted the mouse orthologs and also performed a pairwise local alignment (human-mouse comparison). Further, we performed pairwise local alignment between segments in the coding region which includes INDEL with 30 flanking bases and the similar region in human (human-pig comparison) and mouse (human-mouse comparison). If the sequence identity with human coding regions was at least 50% (human-pig comparison) and the ratio of sequence alignments between INDELs with flanking 30 nucleotide bases in human-pig and human-mouse comparison was at least 0.8, then we considered those regions to be correctly annotated coding regions.

5.4.5 INDEL annotation

INDELs were classified as homopolymer run (HR), if they have six or more identical nucleotide bases adjacent to the INDELs. If flanking region consists of multiple of repeat units identical and directly adjacent to INDELs, these sites are characterized as tandem repeats (TR). INDELs, which are adjacent/surrounded by repetitive regions not identical to INDEL segment, are categorized as predicted hotspot regions (PR). Non-repetitive sites (NR) are regions, which do not fulfill any criteria above i.e., do not have homopolymer runs (HR), do not have multiple adjacent segments similar to the INDELs (TR), and not surrounded by repetitive regions (PR). These NR regions are further classified into two categories, NR-change copy count (NR-CCC) i.e. the presence of INDEL changes local copy count of the segment (max 1 copy), and NR-nonCCC, which do not change the copy count of the segment. The first four classifications of INDELs indicate the DNA polymerase slippage (Montgomery et al. 2013).

5.5 Author's contributions

OM, YP, H-JM, MAMG conceived and designed the experiments. YP, OM performed the experiments and analyzed the data. MAMG RPMAC contributed reagents/materials/analysis tools. YP wrote the manuscript. OM MAMG H-JM RPMAC discussed and improved manuscript. All authors read and approved the final manuscript.

5.6 Acknowledgements

This work was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC Grant agreement no 249894 (SelSweep project). We would like to thank the Swine Genome Consortium for the reference genome build 10.2. We thank Prof. Dr. Ning Li, State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing, China, for providing us DNA samples from Asian wild boars.

5.7 Additional information

Supplementary files and tables can be downloaded from this link: **https://drive.google.com/file/d/0B3goLJGl6JqgOFdUcV84ZVU1aVE/view?usp=sharing**

References

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: Accurate indel calls from short-read data. Genome Res 21: 961–973.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 14: 59–69.

Bosse M, Megens H-J, Madsen O, Paudel Y, Frantz LAF, Schook LB, Crooijmans RPMA, Groenen MAM. 2012. Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. PLoS Genet 8: e1003100.

Chen C, Qiao R, Wei R, Guo Y, Ai H, Ma J, Ren J, Huang L. 2012. A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. BMC Genomics 13: 733.

Choi J-W, Liao X, Park S, Jeon H-J, Chung W-H, Stothard P, Park Y-S, Lee J-K, Lee K-T, Kim S-H, et al. 2013. Massively parallel sequencing of Chikso (Korean brindle cattle) to discover genome-wide SNPs and InDels. Mol Cells 36: 203–211.

Chong Z, Zhai W, Li C, Gao M, Gong Q, Ruan J, Li J, Jiang L, Lv X, Hungate E, et al. 2013. The Evolution of Small Insertions and Deletions in the Coding Genes of Drosophila melanogaster. Mol Biol Evol 30: 2699–2708.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.

Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silio L, Rodriguez M, Groenen M, Ramos-Onsins S, Perez-Enciso M. 2013. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics 14: 148.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. Nucleic Acids Res 40: D84 – D90.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Joshi N, Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human Genome Ultraconserved Elements Are Ultraselected. Science 317: 915–915.

Kerem B, Rommens J, Buchanan J, Markiewicz D, Cox T, Chakravarti A, Buchwald M, Tsui L. 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080.

Kijas JMH, Moller M, Plastow G, Andersson L. 2001. A Frameshift Mutation in MC1R and a High Frequency of Somatic Reversions Cause Black Spotting in Pigs. Genetics 158: 779–785.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078 –2079.

Lugassy J, Itin P, Ishida-Yamamoto A, Holland K, Huson S, Geiger D, Hennies HC, Indelman M, Bercovich D, Uitto J, et al. 2006. Naegeli-Franceschetti-Jadassohn Syndrome and Dermatopathia Pigmentosa Reticularis: Two Allelic Ectodermal Dysplasias Caused by Dominant Mutations in KRT14. Am J Hum Genet 79: 724–730.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16: 1182 – 1190.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470: 59 – 65.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome Res 23: 749–761.

Mullaney JM, Mills RE, Pittard WS, Devine SE. 2010. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet 19: R131–R136.

Neuman JA, Isakov O, Shomron N. 2013. Analysis of insertion–deletion from deep-sequencing data: software evaluation for optimal detection. Brief Bioinform 14: 46–55.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, et al. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411: 603–606.

Paudel Y, Madsen O, Megens H-J, Frantz L, Bosse M, Bastiaansen J, Crooijmans R, Groenen M. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14: 449.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA Loss as a Determinant of Genome Size. Science 287: 1060–1062.

Raeder H, Johansson S, Holm PI, Haldorsen IS, Mas E, Sbarra V, Nermoen I, Eide SA, Grevle L, Bjorkhaug L, et al. 2006. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. Nat Genet 38: 54–62.

Ramayo-Caldas Y, Castello A, Pena RN, Alves E, Mercade A, Souza CA. 2010. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. BMC Genomics 11: 593.

Ren J, Mao H, Zhang Z, Xiao S, Ding N, Huang L. 2011. A 6-bp deletion in the TYRP1 gene causes the brown colouration phenotype in Chinese indigenous pigs. Heredity 106: 862–868.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.

Takahashi M, Saitou N. 2012. Identification and Characterization of Lineage-Specific Highly Conserved Noncoding Sequences in Mammalian Genomes. Genome Biol Evol 4: 641–657.

The 1000 genomes project consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

The chimpanzee sequencing and analysis consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.

Tortereau F, Servin B, Frantz L, Megens H-J, Milan D, Rohrer G, Wiedmann R, Beever J, Archibald A, Schook L, et al. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics 13: 586.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871.

6

General discussion

6.1 Introduction

The overall objective of this thesis was to use next generation sequencing (NGS) data to improve our understanding of the evolution of structural variations such as copy number variations (CNVs), insertions and deletions (INDELs) in pigs and their role in the process of pig domestication and speciation. In chapter 2, I described the dynamics of CNVs in pigs in the context of adaptation and domestication while in chapter 3, I focused on the role of CNVs in the putative ongoing process of speciation in the genus *Sus*. In chapter 4, I investigated the effect of selection and gene clustering on the copy number variation in the olfactory receptor gene family, the most complex and largest gene family in pig genome. In chapter 5, I compared short insertions and deletions to SNPs looking for trends of differences in selection, origin, and genome distribution. In this final chapter, I will explore the findings of the four main chapters of this thesis, and put some of the results into context with other studies. I will also discuss both strengths and limitations of the current tools to study structural variations (SVs), and future trend in the detection of SVs using NGS data.

6.2 Copy number variation and its functional implications

Differences in the number of copies of segments of DNA between different individuals (known as copy number variations) are an ample source of genetic variation in many different organisms including humans (Feuk et al. 2006; Freeman et al. 2006; Redon et al. 2006; Sudmant et al. 2010), mice (Graubert et al. 2007; She et al. 2008; Henrichsen et al. 2009), cattle (Bickhart et al. 2012; Hou et al. 2012), dogs (Axelsson et al. 2013; Freedman et al. 2014), flies (Dopman and Hartl 2007; Emerson et al. 2008), maize (Springer et al. 2009), and yeasts (Zhang et al. 2013). These copy number variable regions (CNVRs) play a prominent role in creating new functional genes, altering gene dosage, reshaping gene structure, and/or modifying the regulatory elements that control gene expression and often resulting in phenotype variation (Long 2001; Otto and Yong 2002; Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010; Dennis et al. 2012). Therefore, understanding

the evolution of CNVs is very important in understanding the contribution of CNVs to the phenotypic evolution of organisms (Long 2001; Otto and Yong 2002; Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010). In chapter 2, I have presented a high resolution CNV map of pigs using NGS data, and consistent with other studies in humans, dogs, mice, and cattle (Redon et al. 2006; Alvarez and Akey 2011; Bickhart et al. 2012), we found that these variations comprise a significant part of their genome. In domestic animals the best known examples of traits that are affected by CNVs are the animal exterior and morphological traits like coat-color in pigs, dorsal hair-ridgeback in dogs, late feathering and pea comb in chickens (Giuffra et al. 2002; Salmon Hillbertz et al. 2007; Elferink et al. 2008; Wright et al. 2009). In chapter 2, we confirmed the coat color CNV related phenotype in pigs and found more than 500 genes, which were variable in copy number in different populations of pigs that could play an important role in phenotypic variation of pigs. Some of these genes were found to be involved in pregnancy related phenotypes, among the most important phenotypes selected for in pig breeding. A recent CNV study in Italian Large White pigs suggested that copy number variable genes such as *ZPLD1* are associated with back fat thickness (Schiavo et al. 2014). Similarly, Fernández et al. (2014) suggested that copy number variable genes such as *SCD* and *USP15* in Iberian pigs, are important in determining the quality of Iberian pig products. Thus, genes variable in copy number are important for phenotypic variation, however, to verify their role in specific phenotypes, gene expression analysis of these copy number variable genes in tissues of interest should be carried out.

Gene expression is a key to underpin the genetic contribution of copy number variable genes to different phenotypes of an organism (Stranger 2007; Charrier et al. 2012; Sudmant et al. 2013). Recently, the ease of obtaining high-resolution whole genome gene expression data allowed researchers to perform a systematic study on classification of phenotypes associated with the expression of genes

(Charrier et al. 2012; Li et al. 2012; Sudmant et al. 2013; Wheeler et al. 2013). In the future, for the most relevant copy number variable genes/gene families in pigs, gene expression analysis in tissues of interest should be carried out that will give provide further clues about the genes/gene families that control the phenotypes in pigs. In addition, efforts to combine all SVs with SNPs will give us a comprehensive map of genetic variation in pig genomes. In combination with other information such as QTL, gene expression, and other phenotypic information, such a comprehensive map will facilitate predictive biological approaches to study genotype/phenotype relations in this livestock species (Bickhart and Liu 2014; Daetwyler et al. 2014).

6.3 CNVR reflect the biogeography, domestication and selection history in pigs

Copy number variations between individuals of a species can result in drastic phenotypic differences (section 6.2) and may therefore be subject to natural selection and selection during domestication. Recent studies on dogs have revealed the importance of CNV of the amylase gene in the early process of dog domestication (Alvarez and Akey 2011; Freedman et al. 2014). Similarly, studies in diverse cattle breeds have suggested the importance of CNVs during cattle domestication, health, and production traits (Liu et al. 2010, 2011; Bickhart et al. 2012; Hou et al. 2012; Shin et al. 2014). Thus, studying CNVs in relation to domestication and demographic history is highly relevant in pigs not only because of their rich natural and domestication history but also because pig is one of the most important livestock species.

Pigs have been domesticated independently at least once in Europe and once in Asia around 10,000 years ago (Larson et al. 2005; Megens et al. 2008). After the initial domestication, pig populations in Asia and Europe have experienced very different selection pressures. Because of these different selection pressures, we hypothesized that differences in SVs in the genomes of wild and domestic pigs will

reflect biogeography, domestication, and selection history of pigs. By including different pigs representing the two independent domestications together with individual representatives of their wild ancestors, in chapter 2, we presented a first comprehensive study on the change in pattern of CNVRs during the process of domestication and/or the natural demographic history of pigs. The comparison of the CNVRs revealed a higher diversity and more CNVRs in Asian pig populations than European pig populations (Figure 2.4). This is consistent with a large population size and a more diverse origin of Asian pigs as observed in genetic variation studies based on SNPs and microsatellites (Groenen et al. 2012; Megens et al. 2008). Thus, pig CNVs are valid predictors of both domestication and demographic history.

By simultaneously examining patterns of variation in SNPs and CNVs in individual of the same human populations, it has been shown that the inferences of population structure based on CNVs and SNPs are generally in agreement (Jakobsson et al. 2008; Wang et al. 2010). In contrast, Itsara et al., (2009) suggested limited evidence for stratification of CNVs in geographically distinct human populations. Since, the analyses were based on SNP-arrays, the variability of genotyping intensity across the genome influences the ability of CNV detection tools such as PennCNV to identify CNVs by systematically giving rise to additional false-positive calls (Wang et al. 2007; Itsara et al. 2009). Wang et al. (2010), therefore excluded those high variance samples from the study by Jakobssen et al. (2008), re-analyzed the CNVs and found support for a SNP and CNV based inference of human population structure. In addition, the studies in human were carried out using SNP-arrays (Jakobsson et al. 2008; Itsara et al. 2009; Wang et al. 2010) so, it is not clear how much ascertainment bias and exclusion of troublesome SNPs (located in CNVRs) from the SNP arrays affect the results by introducing false negative CNV calls in the ascertained populations. In the CNV study presented in chapter 2, we have used NGS data, which is able to detect most of CNVRs in the genome and is independent

from SNPs; however, we also missed small CNVRs in the genome (<6Kb), which might have resulted in an underestimation of the level of diversity of CNVs in the different pig populations. Thus, to have a better estimation of the effect of demography on CNVs in pigs, the CNV patterns and SNPs across diverse populations in a much larger cohort of samples should be investigated.

In previous analyses, we observed a deep phylogenetic split between European and Asian pigs (Groenen et al. 2012; Frantz et al. 2013). Thus, the sequence divergence of Asian and European pigs and the European origin of the reference genome might have resulted in an underestimation of *de novo* CNVRs specific to Asian populations. However, we observed higher diversity among the Asian wild and domestic populations (chapter 2), which suggests a minor impact of the reference, at least to detect CNVRs between different populations. One solution to resolve such ascertainment bias would be to have a separate reference genome for all studied populations. Recently, *de novo* assembled references of some other pigs from Asia and Europe have become available (Fang et al. 2012; Li et al. 2013; Vamathevan et al. 2013). Even though we need extra care to implement these assembled genomes for the analysis of CNVs (see below), future studies generating comprehensive maps of CNVRs should carefully include some other reference genomes of pigs from different populations to avoid reference biased ascertainment. Since human populations are not as variable as pigs, the inclusion of other references is especially important for pigs.

6.4 Copy number variable genes/gene families in pigs

Several studies in different organisms have found genes involved in environmental response to be over-represented in CNVRs (Sebat et al. 2004; Tuzun et al. 2005; Redon et al. 2006; She et al. 2008; Nicholas et al. 2009; Alvarez and Akey 2011; Bickhart et al. 2012). Our gene enrichment analysis also suggested that the majority of the genes and gene families overlapping CNVRs in pigs are involved in biological

processes regulating sensory perception of smell, signal transduction, neurological processes, and metabolic processes (chapter 2). However, due to the size limitation of our CNVRs (≥6Kb), genes residing in CNVRs smaller than 6Kb were not identified and our list of copy number variable genes is therefore not complete. The next logical step was to conduct a comprehensive gene level CNV analyses. In chapter 4, we present a novel approach to identify the CNV status of genes. We concentrated our study on one of the largest and complex gene families in the pig genome i.e. the olfactory receptor (OR) gene family because these genes appeared to be highly copy number variable in pigs (chapter 2 and 3) and because they have a simple gene structure (i.e. only one open reading frame), which facilitated the analysis. Furthermore, OR genes are generally found in clusters in the genome, which enabled us to look at the effect of gene clustering in the evolution of the CNV of genes. This gene level CNV study (chapter 4) showed that we had underestimated the CNV of ORs in the pig OR repertoire in our previous study (chapter 2). In addition, we found that both selection and clustering play a role in the variation of ORs in the pig genome (chapter 4). A recent CNV study of ORs in the human genome suggested the role of genetic drift in the variation of ORs between different human populations (Nozawa et al. 2007). In contrast, other studies in human suggested a role for selective constraints and CNV formation biases in the variation of ORs of human OR repertoire (Hasin et al. 2008; Young et al. 2008). However, these studies in human are based on large CNVRs and thus lack information of the copy number variable status of some of the ORs in the human genome, which might have influenced the result. With the new pipeline described in chapter 4, we overcame the issue of underestimating the CNV of ORs in the OR repertoire. The cross alignment, i.e. off target mapping of short sequence reads, between highly similar ORs (some are 100% identical) prevented estimation of absolute copy number of ORs. That restrained our further analysis on the expansion and contraction of the OR families and subfamilies in different populations. The problem of cross alignment could be resolved once sequence data

with longer fragment length are obtained with e.g. sequencing technology like PacBio**,** Illumina TruSeq Synthetic Long-Reads and Pseudo-Sanger sequence (Branton et al. 2008; Eid et al. 2009; Ruan et al. 2013; McCoy et al. 2014). The longer sequence reads will allow the assembly and re-construction of all the copies of the ORs. This will enable deciphering families/members of ORs expanding in one population and diminishing in another population. In addition, it allows the identification of the mechanisms behind expansion and/or contraction of specific families/members of ORs between populations. However as described above (section 6.2), before making any conclusions, the real effect of expanding and contracting gene families must be tested by gene expression analyses.

6.5 Copy number variable regions and pig speciation

Speciation is the fundamental evolutionary process that drives ecological diversification on earth (Mayr 1963; Mallet 1995; Coyne and Orr 2004). There are many different models for modes of speciation in nature (chapter 3) and has been hypothesized that only a few sporadic genetic changes are needed to promote evolution of a new species (Ellegren et al. 2012; Martin et al. 2013). Some speciation genes have been identified (Mihola et al. 2009; Perez and Wu 1995; Masly et al. 2006), but from these few genes, it is hard to suggest any common or general pathway leading to speciation. Recent studies on butterfly and flycatcher genomes suggest that speciation is driven by certain regions in the genome (also known as islands of speciation) that promote divergence between highly related populations even when these populations occupy a similar geographical location (Ellegren et al. 2012; Martin et al. 2013). Genes in those islands of divergence were found to be involved in meiosis and the production of gender cells. In addition, the separation of these sub-species also seems to be caused by the dissimilarities in chromosome structures, which make recombination between divergent haplotypes impossible. Thus, chromosome rearrangement and other structural variations, rather than different adaptations of individual genes can cause speciation. In

chapter 3, we investigated the role of CNVRs during the ongoing speciation in the genus *Sus*. We observed a faster rate of evolution of CNVRs compared to single nucleotide polymorphism (chapter 3). Since fast evolving regions potentially play a role in the transition from pre- to post-zygotic isolation, the observed elevated evolutionary rate of CNVs suggested that some of the CNVRs could be involved in speciation. The fact that many of these CNVRs overlap with ORs and that ORs play an important role in food foraging and finding potential mates in pigs (Groenen et al. 2012; Nguyen et al. 2012), led us to test the role of ORs in pig speciation. Compared to other CNVRs in the genome, CNVRs overlapping ORs recapitulated the well-accepted phylogeny of the genus *Sus*, whereas the CNVRs, which did not overlap ORs, demonstrated evidence of admixture and/or genetic drift. This supported our hypothesis that the CNVRs overlapping ORs acted as medium to adapt to different environments and may be involved in preventing different subspecies of genus *Sus* to be admixed, thereby triggering the process of diversification.

To understand the importance of OR subfamilies in different ecological niche, the functional OR sub-genome across all the subspecies of genus *Sus* should be compared using a phylogenetic analysis. Moreover, we also need detailed information of odorants that bind to each OR. Although, in chapter 3, we uncovered some examples of expanding and contracting OR families/subfamilies between different species of the genus *Sus*, due to the limitation of our computational approach, we could not assess whether the expanded OR families/subfamilies are still functional. As described above (section 6.2 and 6.4), longer sequence reads and gene expression analyses of relevant tissues would facilitate the estimation of the functional status of the expanded/contracted OR families. Future studies should focus on constructing detailed maps of the OR in OR repertoire for all the species of the genus *Sus*, including information of expanding and contracting OR families in each species. Such maps may provide further insight

into the role of selection in the copy number of ORs for the different species. For instance, phylogenetic analyses, and other statistical methods could be used to partition these ORs based on the ecotype. Further, functional studies can aid in the identification of families of ORs that might have played a role in adaptation during parapatric periods and mate selection during sympatric periods that could have driven divergence and ultimately speciation in genus *Sus*.

6.6 Limitations and improvements in the current technology

6.6.1 Tools to detect SVs

Structural variations are a major source of intra and inter-specific genotypic variation and have been shown to play an important role in genome evolution (Hurles et al. 2008; Stankiewicz and Lupski 2010; Gokcumen et al. 2013). Thus, computational biologists are actively developing and improving tools to process genomes efficiently to identify SVs. SAMTools (Li et al. 2009), GATK (DePristo et al. 2011), BreakDancer (Chen et al. 2009), Pindel (Ye et al. 2009), and Dindel (Albers et al. 2011), are only a few examples of tools that can detect SVs. The increasing number of SV detection tools, also, makes it complicated to compare the relative performance of tools and select the one best suited for a given set of data. The majority of SV detection tools have been developed to include different information sources such as read depth, discordant paired-end reads, split-read alignments, and assembled segments (chapter 1). The SV detection tools are also specialized to detect certain types and sizes of SVs, using different types of NGS data from specific platforms and optimized for certain species. Similar to what is the case for SNP callers, none of the SV detection tools are in complete agreement in SV calls. The agreement between SNP callers is around 60% whereas the agreement between SV callers is around 43% (Alkan et al. 2011a; O'Rawe et al. 2013). Another problem is that every method publishes its own simulation-based evaluation datasets, while the documentation about pre- and post-processing are

not always clear enough to reproduce and interpret the results (Wong et al. 2010; Mimori et al. 2013; Leung et al. 2014). Thus, a future challenge for researchers working on SVs is to make a good selection of tools, which generate a set of SV calls that is both comprehensive and reliable. Systematic benchmarking and evaluation of SV detection tools are therefore essential to measure the accuracy, performance, and robustness of tools, and to improve them (Wong et al. 2010; Mimori et al. 2013; Leung et al. 2014). Because of the lack of an experimentally validated standard dataset, the benchmarking efforts are also error prone. Thus, the establishment of standard datasets for phylogenetically diverse organisms to benchmark these tools will be the first and best strategy, which not only eases method comparisons but also guides further improvement of SV detection tools (Mimori et al. 2013; Leung et al. 2014).

6.6.2 The reference genome

Another major problem in determining reliable SV is the wide range of in-completeness of reference genomes. A complete reference genome sequence of a species includes all the features of a genome such as genes, repetitive elements, and regulatory elements. A reference genome of an organism aids the identification and interpretation of genomic variations of the organism by allowing reads to align/assemble, which significantly reduces computational load involved in an analysis. Alignment of sequencing reads against a reference genome, therefore, is the first major data processing step in genome research involving NGS data. To obtain a comprehensive and reliable set of information of genomic variations, the reference genome therefore, needs to be as complete and error free as possible. All the methods, especially those based on NGS data described in this thesis, heavily rely on the quality of the reference genome for correctly detecting and interpreting results. Many species, including livestock species such as cattle, pig, goat, chicken, sheep have high quality draft reference genomes available (Bovine Genome Sequencing and Analysis Consortium 2009; Groenen et al. 2012; Dong et

al. 2013; Jiang et al. 2014) and these reference genomes are being used to detect all types of genomic variations including SVs. However, incompleteness and errors in the assembled reference genomes can significantly frustrate prediction of SVs and results are often misleading or difficult to interpret. For example, due to the large number of gaps present in the pig reference genome, in the studies described in chapters 2 and 3, we were limited to CNVRs smaller than 98 Kb. Similarly, our unpublished analysis based on data from the reference individual itself (TJ Tabasco) for e.g. inversions and translocations revealed a very high level of noise suggesting a considerable level of assembly errors in the current genome build making it very difficult to perform such SV analysis. There are some examples of misleading reports which were caused by errors in the reference genome such as a cattle segmental duplication of 39 Mb that was due to the result of an assembly error (build Btau4.2 (Zimin et al. 2009)). Likewise, more than 14 Mb of segmental duplications in the chimpanzee and chicken genomes are in reality caused by assembly errors (Kelley and Salzberg 2010). Thus, working with an incomplete draft reference genome, requires caution when interpreting the results. To avoid misinterpretation of results due to problems in assembly, consortiums, which are involved in developing reference genomes, should also try to speed the process of improving the reference genome by integrating unassembled contigs/scaffolds and if possible filling gaps present in the draft reference genome.

6.6.3 Phenotypic information

The term used to describe the observable characteristics of an organism is called phenotype, whereas the term used to denote genetic make-up of an organism is called genotype. The genotype functions as a set of instruction for the phenotype. The phenotypic variations between species are the result from variations in their genotype, environment (e.g. diet, climate, illness, and stress), and genome modification (epigenetics). Advances in sequencing technology now make it possible to sequence entire genomes of many individuals/species and study genetic

variations in a short time. However, we still know very little about the precise relationships between genotypic and phenotypic variation. Linking the genotype to phenotype is, therefore, one of the most challenging aspects of contemporary genome research. Due to the incomplete annotation of the pig reference genome and the very limited information available on phenotypes of the pigs used in this thesis, it is almost impossible to link the detected SVs to certain phenotypes. Hence, more detailed and extensive phenotypic information from large cohorts of animals is vital to interpret and understand the role of SVs on different phenotypes. Besides phenotypic data it is also important to include environmental factors, which might influence the phenotype of an individual when analyzing genotype/phenotype interactions (Smith-Tsurkan et al. 2013; Corrigan et al. 2011; Fischer et al. 2004).

6.7 Future trends

Different methods have been developed to use different types of signals/signatures obtained from the NGS data to detect SVs (chapter 1). However, these methods are mostly based on mapping sequences against a reference genome. At the moment, there is not any comprehensive algorithm or solution that can be used to identify all SVs present in a genome irrespective of their class, type, and frequency. Due to the higher rate of false calling of SVs, researchers are using multiple approaches/tools to identify and verify their results, which is costly both in time and money. Thus, different groups involved in SV detection are working to develop algorithms that can use all types of signals obtained from the NGS data to detect all SVs in a genome. An approach different groups are interested in, is to develop a "mixed/hybrid approach". The mixed/hybrid approach combines available multiple approaches described in chapter 1, for example, considering a paired-end approach in the context of split reads and/or read depth and/or an assembly (Ye et al. 2009; Hajirasouliha et al. 2010; Hormozdiari et al. 2010; Medvedev et al. 2010; McKenna et al. 2010; Zeitouni et al. 2010; Albers et al. 2011; Rausch et al. 2012; Marschall et

al. 2012). Other groups are working on benchmarking the available tools to find the best tool and/or best approach to merge the results from multiple tools to obtain better SV calls (Mimori et al. 2013; Leung et al. 2014). However, an incomplete reference genome will always be a limitation for these approaches as these approaches rely on alignment of sequences against the reference genome. A distinct approach to call variants, which basically overcomes all the problems of sequence alignment based approaches, is to assemble sequence reads into contigs/scaffolds and compare the assemblies of newly sequenced individuals to discover variants (Li et al. 2011). Current *de novo* assembly algorithms (chapter 1) have substantial computational requirements and require sequencing of fragments with varying insert sizes. Multiple plant, bird, and mammalian genomes have so far been *de novo* assembled solely using sequence data generated using NGS platforms (Li et al. 2010; Al-Dous et al. 2011; Fang et al. 2012; Wang et al. 2012; Li et al. 2013; Vamathevan et al. 2013; Ganapathy et al. 2014). However, these *de novo* assembled genomes are constructed based on approaches that assume that similar sequence reads originated from the same genomic region allowing overlapping reads to be merged to reconstruct the underlying genome sequence (Nagarajan and Pop 2013). The assumption is valid only when there are no repetitive regions in the genome. That assumption is not realistic as in many Eukaryotes the proportion of the total genome containing repeats is substantial, and for instance in mammals this typically is more than 40% (Feschotte et al. 2002; de Koning et al. 2011; McCoy et al. 2014). This complicates the assembly and may induce assembly failure. Recently developed technology such as pseudo-sanger sequencing, which claims to use paired-end NGS data to fill gaps between paired-ends to generate error-free longer sequences equivalent to the conventional Sanger reads in length, could be beneficial for *de novo* assembly algorithms specially to resolve complex regions in the genome (Ruan et al. 2013). Similarly, longer reads generated by PacBio, Illumina TruSeq Synthetic Long-Reads, Nanopore could also resolve such issues related to repeats and other complicated region in the genome and improve *de*

novo assembly (Clarke et al. 2009; English et al. 2012; McCoy et al. 2014). However, even after including fragments of different read length and insert sizes, a high degree of variability between different *de novo* assembly algorithms was observed (Alkan et al. 2011b; Ye et al. 2011; Bradnam et al. 2013). Therefore, for now the best solution would be to have multiple libraries with varying insert sizes and read lengths and to incorporate both alignment and assembly based "assembly-alignment-hybrid" approaches which empower comparative analyses that will facilitate researchers to identify all types of SVs present in the genome.

References

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: Accurate indel calls from short-read data. Genome Res 21: 961–973.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al. 2011. De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat Biotech 29: 521–527.

Alkan C, Coe BP, Eichler EE. 2011a. Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363–376.

Alkan C, Sajjadian S, Eichler EE. 2011b. Limitations of next-generation genome sequence assembly. Nat Meth 8: 61–65.

Alvarez C, Akey J. 2011. Copy number variation in the domestic dog. Mamm Genome 1–20.

Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495: 360–364.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22: 778 – 790.

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. Front Genet 5. http://www.frontiersin.org/Journal/Abstract.aspx?s=1254&name=evolutionary_and_population_genetics&ART_DOI=10.3389/fgene.2014.00037.

Bovine Genome Sequencing and Analysis Consortium. 2009. The Genome Sequence of Taurine Cattle: a window to ruminant biology and evolution. Science 324: 522 – 528.

Bradnam K, Fass J, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman J, Chapuis G, Chikhi R, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2: 10.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al. 2008. The potential and challenges of nanopore sequencing. Nat Biotech 26: 1146–1153.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny during Spine Maturation. Cell 149: 923–935.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Meth 6: 677–681.

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nano 4: 265–270.

Corrigan LJ, Lucas MC, Winfield IJ, Hoelzel AR. 2011. Environmental factors associated with genetic and phenotypic divergence among sympatric populations of Arctic charr (Salvelinus alpinus). J Evol Biol 24: 1906–1917.

Coyne JA, Orr HA. 2004. Speciation. Sinauer Associates Sunderland, MA.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet advance online publication. http://dx.doi.org/10.1038/ng.3034.

De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS Genet 7: e1002384.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell 149: 912–922.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.

Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J, et al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). Nat Biotech 31: 135–141.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in Drosophila melanogaster. Proc Natl Acad Sci 104: 19920–19925.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323: 133–138.

Elferink M, Vallee A, Jungerius A, Crooijmans R, Groenen M. 2008. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. BMC Genomics 9: 391.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491: 756–760.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320: 1629–1631.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7: e47768.

Fang X, Mou Y, Huang Z, Li Y, Han L, Zhang Y, Feng Y, Chen Y, Jiang X, Zhao W, et al. 2012. The sequence and analysis of a Chinese pig genome. GigaScience 1: 16.

Fernández AI, Barragán C, Fernández A, Rodríguez MC, Villanueva B. 2014. Copy number variants in a highly inbred Iberian porcine strain. Anim Genet 45: 357–366.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3: 329–341.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7: 85 – 97.

Fischer K, Bot ANM, Zwaan BJ, Brakefield PM. 2004. Genetic and environmental sources of egg size variation in the butterfly Bicyclus anynana. Heredity 92: 163–169.

Frantz L, Schraiber J, Madsen O, Megens H-J, Bosse M, Paudel Y, Semiadi G, Meijaard E, Li N, Crooijmans R, et al. 2013. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biol 14: R107.

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genet 10: e1004016.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al. 2006. Copy number variation: New insights in genome diversity. Genome Res 16: 949–961.

Ganapathy G, Howard J, Ward J, Li J, Li B, Li Y, Xiong Y, Zhang Y, Zhou S, Schwartz D, et al. 2014. High-coverage sequencing and annotated assemblies of the budgerigar genome. GigaScience 3: 11.

Giuffra E, Tornsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, Anderson SI, Archibald AL, Andersson L. 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. Mamm Genome 13: 569 – 577.

Gokcumen O, Tischler V, Tica J, Zhu Q, Iskow RC, Lee E, Fritz MH-Y, Langdon A, Stütz AM, Pavlidis P, et al. 2013. Primate genome architecture influences structural

variation mechanisms and functional consequences. Proc Natl Acad Sci 110: 15764–15769.

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, et al. 2007. A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet 3: e3.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491: 393–398.

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC. 2010. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26: 1277–1283.

Hasin Y, Olender T, Khen M, Gonzaga-Jauregui C, Kim PM, Urban AE, Snyder M, Gerstein MB, Lancet D, Korbel JO. 2008. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. PLoS Genet 4: e1000249.

Henrichsen CN, Chaignat E, Reymond A. 2009. Copy number variants, diseases and gene expression. Hum Mol Genet 18: R1–R8.

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26: i350–i357.

Hou Y, Liu G, Bickhart D, Matukumalli L, Li C, Song J, Gasbarre L, Tassell C, Sonstegard T. 2012. Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. Funct Integr Genomics 12: 81–92.

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. Trends Genet 24: 238–245.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97–108.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, et al. 2009. Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. Am J Hum Genet 84: 148–161.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.

Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W, et al. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344: 1168–1173.

Kelley D, Salzberg S. 2010. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol 11: R28.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. Spec Issue Mem John Maynard Smith Spec Issue Mem John Maynard Smith 239: 141–151.

Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E, et al. 2005. Worldwide Phylogeography of Wild Boar Reveals Multiple Centers of Pig Domestication. Science 307: 1618 –1621.

Leung WY, Marschall T, Paudel Y, Falquet L, Mei H, Schoenhuth A, Maoz TY. 2014. The new face of SV detection and tool development for all species SV-AUTOPILOT: Structural Variation AUTOmated PIpeLine Optimization Tool. Submitted.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078 –2079.

Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, Wang T, Yeung CKL, Chen L, Ma J, et al. 2013. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet 45: 1431–1438.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010. The sequence and de novo assembly of the giant panda genome. Nature 463: 311–317.

Liu GE, Brown T, Hebert DA, Cardone MF, Hou Y, Choudhary RK, Shaffer J, Amazu C, Connor EE, Ventura M, et al. 2011. Initial analysis of copy number variations in cattle selected for resistance or susceptibility to intestinal nematodes. Mamm Genome 22: 111 – 121.

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. Genome Res 39: 693 – 703.

Li Y, Shaw CA, Sheffer I, Sule N, Powell SZ, Dawson B, Zaidi SNY, Bucasas KL, Lupski JR, Wilhelmsen KC, et al. 2012. Integrated copy number and gene expression analysis detects a CREB1 association with Alzheimer/'s disease. Transl Psychiatry 2: e192.

Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H, et al. 2011. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat Biotech 29: 723–730.

Long M. 2001. Evolution of novel genes. Curr Opin Genet Dev 11: 673–680.

Mallet J. 1995. A species definition for the modern synthesis. Trends Ecol Evol 10: 294–299.

Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, Schliep A, Schönhuth A. 2012. CLEVER: clique-enumerating variant finder. Bioinformatics 28: 2875–2882.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res 23: 1817–1828.

Masly JP, Jones CD, Noor MAF, Locke J, Orr HA. 2006. Gene Transposition as a Cause of Hybrid Sterility in Drosophila. Science 313: 1448–1450.

Mayr E. 1963. Animal species and evolution. Anim Species Their Evol.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier A-S. 2014. Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PLoS ONE 9: e106689.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. 2010. Detecting copy number variation with mated short reads. Genome Res 20: 1613–1622.

Megens HJ, Crooijmans R, San Cristobal M, Hui X, Li N, Groenen MA. 2008. Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genet Sel Evol 40: 103 – 128.

Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. 2009. A Mouse Speciation Gene Encodes a Meiotic Histone H3 Methyltransferase. Science 323: 373–375.

Mimori T, Nariai N, Kojima K, Takahashi M, Ono A, Sato Y, Yamaguchi-Kabata Y, Nagasaki M. 2013. iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. BMC Syst Biol 7: S8.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. Nat Rev Genet 14: 157–167.

Nguyen D, Lee K, Choi H, Choi M, Le M, Song N, Kim J-H, Seo H, Oh J-W, Lee K, et al. 2012. The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. BMC Genomics 13: 584.

Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. 2009. The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Res 39: 491 – 499.

Nozawa M, Kawahara Y, Nei M. 2007. Genomic drift and copy number variation of sensory receptor genes in humans. Proc Natl Acad Sci 104: 20421–20426.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5: 28.

Otto SP, Yong P. 2002. The evolution of gene duplicates. In Advances in Genetics (ed. Jay C. Dunlap and C.-ting Wu), Vol. Volume 46 of, pp. 451–483, Academic Press.

Perez DE, Wu CI. 1995. Further characterization of the Odysseus locus of hybrid sterility in Drosophila: one gene is not enough. Genetics 140: 201–206.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28: i333–i339.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. Nature 444: 444 – 454.

Ruan J, Jiang L, Chong Z, Gong Q, Li H, Li C, Tao Y, Zheng C, Zhai W, Turissini D, et al. 2013. Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology. BMC Genomics 14: 711.

Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, Wade CM, von Euler H, Gustafson U, Hedhammar A, et al. 2007. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat Genet 39: 1318 – 1320.

Schiavo G, Dolezal MA, Scotti E, Bertolini F, Calò DG, Galimberti G, Russo V, Fontanesi L. 2014. Copy number variants in Italian Large White pigs detected using high-density single nucleotide polymorphisms and their association with back fat thickness. Anim Genet n/a–n/a.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. 2004. Large-Scale Copy Number Polymorphism in the Human Genome. Science 305: 525 –528.

She X, Cheng Z, Zollner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. Nat Genet 40: 909–914.

Shin D-H, Lee H-J, Cho S, Kim H, Hwang J, Lee C-K, Jeong J, Yoon D, Kim H. 2014. Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level. BMC Genomics 15: 240.

Smith-Tsurkan SD, Herr RA, Khuder S, Wilke CO, Novella IS. 2013. The role of environmental factors on the evolution of phenotypic diversity in vesicular stomatitis virus populations. J Gen Virol 94: 860–868.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al. 2009. Maize Inbreds Exhibit High Levels of Copy Number

Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet 5: e1000734.

Stankiewicz P, Lupski JR. 2010. Structural Variation in the Human Genome and its Role in Disease. Annu Rev Med 61: 437–455.

Stranger BE. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23: 1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Project 1000 Genomes, et al. 2010. Diversity of human copy number variation and multicopy genes. Science 39: 641 – 646.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. Nat Genet 37: 727–732.

Vamathevan JJ, Hall MD, Hasan S, Woollard PM, Xu M, Yang Y, Li X, Wang X, Kenny S, Brown JR, et al. 2013. Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. Toxicol Appl Pharmacol 270: 149–157.

Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA. 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat Appl Genet Mol Biol 9.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665 – 1674.

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, et al. 2012. The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. Plant J 72: 461–473.

Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, Garg S, Henning E, Blackburn H, Loos RJF, Wareham NJ, et al. 2013. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat Genet 45: 513–517.

Wong K, Keane T, Stalker J, Adams D. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol 11: R128.

Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, Vieaud A, Tixier-Boichard M, Rubin C-J, Imsland F, Hallböök F, et al. 2009. Copy Number Variation

in Intron 1 of SOX5 Causes the Pea-comb Phenotype in Chickens. PLoS Genet 5: e1000512.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871.

Ye L, Hillier L, Minx P, Thane N, Locke D, Martin J, Chen L, Mitreva M, Miller J, Haub K, et al. 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol 12: R31.

Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. 2008. Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family. Am J Hum Genet 83: 228–242.

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E. 2010. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26: 1895–1896.

Zhang H, Zeidler AFB, Song W, Puccia CM, Malc E, Greenwell PW, Mieczkowski PA, Petes TD, Argueso JL. 2013. Gene Copy-Number Variation in Haploid and Diploid Strains of the Yeast Saccharomyces cerevisiae. Genetics 193: 785–801.

Zimin A, Delcher A, Florea L, Kelley D, Schatz M, Puiu D, Hanrahan F, Pertea G, Van Tassell C, Sonstegard T, et al. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10: R42.

S

Summary

Summary

Structural variations (SVs) are chromosomal rearrangements in a genome such as insertions, deletions, inversions, translocations, and copy number variations (CNVs). These variations contribute to genomic variation and may influence the phenotypes of organisms. It has been shown that SVs are as important as single nucleotide polymorphisms (SNPs) in phenotypic variation and involve more base differences between individuals than SNPs. The overall objective of this thesis was to use next generation sequencing (NGS) data to improve our understanding of the evolution of SVs in pigs and their role in the process of domestication and speciation. In chapter 1, I described different mechanisms that generate SVs in a genome, the implication of SVs in diseases and phenotypic traits, and the impact of SVs on genome evolution. Furthermore, I also described the different approaches to detect SVs in a genome.

There are very few studies where NGS data have been used to understand the dynamics of SVs during the process of domestication and speciation. Generally, the absence of ancestral wild populations and the lack of proper samples from different biogeographic regions have made it difficult to unravel questions related to the impact of SVs on domestication processes. In that respect, pigs are a very good model to perform such studies as pigs were domesticated several times, independently, from local wild populations in Asia and Europe. Due to the extensive selective pressures, the present day pig breeds from Asia and Europe have very distinct phenotypic characteristics compared to their wild counterpart. Thus, analysis of SVs in genomes of wild and domestic pig populations from different Eurasian regions provide an excellent opportunity to assess the evolution of SVs in relation to the domestication, selection and the biogeographic history of pigs. In chapter 2, we investigated copy number variable regions (CNVRs) in 16 different individual pigs from different wild and domestic populations from Asia and Europe by whole genome re-sequencing. We identified 3,118 CNVRs overlapping genes related to e.g. sensory perception, neurological process, and

response to stimulus. The majority of CNVRs ascertained in domestic pigs are also observed to be variable in wild boars and very few CNVRs seem to have been specifically under selection during domestication. This suggests that the vast majority of CNVRs has not been involved in the phenotypic differences between wild and domestic pigs. Instead, the majority of the variable regions were found to reflect the demographic pattern of pigs, which is in line with other types of variations such as SNPs and microsatellites. Our study represents a comprehensive analysis of CNVR in both domestic and wild pigs and provides valuable insights in the evolutionary dynamics of CNV in the context of adaptation and domestication.

Different studies suggested that SVs, especially CNVs that involve gene duplication and deletion, could be a predominant mechanism driving gene and genome evolution. Studies in primates have provided detailed knowledge on the potential evolutionary roles of CNVRs between species. However, our understanding of their significance during the ongoing process of speciation is hindered by the lack of CNV data from evolutionarily closely related species where speciation is still ongoing. The pig species of the genus *Sus* from Island of South East Asia, i.e. Java, Borneo, Sulawesi, and the Philippines is well suited to study the process of speciation since these morphologically well defined species are still capable of producing fertile offspring, suggesting that the process of speciation is still ongoing. We sequenced the genomes and mapped CNVs in these five closely related species of the genus *Sus* to investigate the role of CNVs in speciation. We identified 1,408 CNVRs and observed a highly significant overrepresentation of olfactory receptor genes (ORs) in those CNVRs. Different phylogenetic analyses based on CNVRs that overlap ORs supported the well-accepted topology of the genus *Sus*, whereas phylogenetic analysis on all CNVRs and CNVRs overlapping genes other than ORs showed evidence of random drift and/or admixture. We therefore hypothesize that inter-specific variation in copy number of ORs provided the means for rapid adaptation to different environments during the diversification of the genus *Sus* in the

Pliocene. Furthermore, these regions might have acted as barriers preventing massive gene flow between these species during the multiple hybridization events that took place later in the Pleistocene, thus suggesting a possible role of ORs in pig speciation.

In Chapter 2 and 3, we focused on CNVRs in the genome larger than 5Kb thus lacking the variation status of genes and other functionally important regions in the genome that are smaller than 5Kb. Therefore, the next natural step for us was to conduct a comprehensive CNV study on some of these regions. In chapter 4, we developed and discussed a novel approach to identify variation on the gene level. We focused on the OR gene family due to the structure, organization and overrepresentation of ORs in the CNVRs in pigs (chapter 2 and 3). The gene-based approach outperforms other approaches to accurately predict CNV of each gene in the OR repertoire. We further investigated the significance of selection and genetic drift in the evolution of ORs in the pig genome by sequencing 36 wild and domesticated pigs from Asia and Europe. Around 60% of ORs were found to vary in copy number. Most of the copy number variable ORs reside in clusters, suggesting an important role of gene clusters in promoting the variation of copy number through non-allelic homologous recombination (NAHR). The higher degrees of intra- and inter-population divergence of functional ORs indicate a probable role of selection on the variation of functional ORs in the pig genome. Surprisingly the distribution of the relative copy number of non-functional ORs is significantly different from a normal distribution as expected by neutral evolution of non-functional ORs. Since, both functional and non-functional ORs reside in clusters in the genome, NAHR might simultaneously have facilitated the variation of both functional and non-functional ORs resulting in high variation of both functional and non-functional ORs. Thus, we conclude that both selection and genome clustering of ORs in the genome play important roles in overall copy number variation of the OR repertoire in pigs.

In chapter 5, we used whole genome sequences of 42 wild and domestic pigs from four different populations from Europe and Asia to generate a detailed map of short insertions and deletions (INDELs) and single nucleotide polymorphisms (SNPs). We reported over 0.5 million INDELs (size ≤10 bases) and over 6.0 million SNPs per population. The INDELs and SNPs are distributed throughout the pig genome and the U-shaped distributions suggest a higher density of INDELs and SNPs at the ends of the chromosomes. We found polymerase slippage as the major mechanism of INDEL formation in the pig genome. On average, more than 165,000 INDELs per population were mapped to annotated pig genes, with on average 422 coding regions affected by INDELs resulting in a frame shift. However, gene expression analysis suggested that most of the coding regions with INDELs that result in a frame shift were found erroneously annotated as coding region, supporting strong purifying selection against INDELs in functionally important regions. Thus, comprehensive verification of annotation integrity should be applied before deciphering possible involvement of INDELs in disease and other traits in the pig genome.

Finally, in chapter 6, I discussed the main findings of this thesis regarding CNVs and their importance in domestication and speciation. In addition, I also discussed strengths and limitations of current tools and future trends in the detection of SVs using NGS data.

S

Samenvatting

Samenvatting

Structurele variaties zijn chromosomale veranderingen in een genoom, zoals inserties, deleties, inversies, translocaties, en variatie in het aantal kopieën (Copy Number Variation, of CNV). Deze structurele variaties (SV) dragen bij aan de genoom-wijde variatie en kunnen in potentie het fenotype van het organisme beïnvloeden. Het is aangetoond dat SVs minstens even belangrijk zijn als variatie in een enkele nucleotide (Single Nucleotide Polymorphism, of SNP) voor het verklaren van fenotypische variatie. Gekwantificeerd als het aantal basenparen dat verschilt tussen individuen van dezelfde soort vormen CNVs een grotere bron van variatie dan SNPs.

Het doel van dit proefschrift was om middels de nieuwste technologie om DNA basenvolgorde te bepalen (Next Generation Sequencing, of NGS) ons begrip van SVs in het varken te vergroten, met name in het licht van de evolutie en domesticatie van deze soort.

In hoofdstuk 1 beschrijf ik de verschillende mechanismen die SVs in het genoom kunnen veroorzaken, de relatie tussen SVs en ziektes en andere kenmerken, en de relatie tussen SVs en genoomevolutie. In dit hoofdstuk worden ook de diverse methoden om SVs in het genoom te karakteriseren beschreven.

Tot op heden zijn er slechts weinig studies gepubliceerd waarin NGS data wordt gebruikt om de dynamiek van SVs tijdens het proces van domesticatie en soortvorming te begrijpen. De afwezigheid van de wilde, voorouderlijke vorm van veel van onze landbouwhuisdieren, in elk geval in die gebieden waar domesticatie voor het eerst plaatsvond, is een belangrijke beperking in genetische studies naar domesticatie. Het varken is in dat opzicht een uitzondering en daardoor een uitstekend model voor domesticatie. Het varken is meerdere keren, onafhankelijk van elkaar, gedomesticeerd in Europa en Azië, vanuit het – nog steeds - ter plekke voorkomende wilde zwijn. Vanwege selectie op allerlei verschillende kenmerken

verschillen de huidige Europese en Aziatische varkensrassen in hoge mate van de huidige wilde zwijnen.

Het varken is dus bij uitstek geschikt om de rol van SVs in domesticatie en selectie te bestuderen en om de biogeografische geschiedenis van deze landbouwhuisdieren te beschrijven. In hoofdstuk 2 worden gebieden in het genoom die CNVs bevatten (Copy Number Variable Regions, of CNVR) in 16 verschillende varkens, die verschillende wilde en gedomesticeerde populaties uit Azië en Europa vertegenwoordigen, bestudeerd met behulp van NGS technieken. In totaal werden 3.118 CNVRs ontdekt. Deze gebieden overlappen met genen die betrokken zijn in zintuiglijke waarneming, neurologische processen, en stimulusrespons. Het grootste deel van de genoomregio's met CNVRs in het varken bevatten ook CNVs in wilde zwijnen, en slechts weinig van de CNVRs lijken een hoge mate van selectie te hebben ondervonden gedurende het domesticatieproces. Dit suggereert dat het grootste deel van de CNVRs geen directe relatie heeft met de fenotypische verschillen tussen wilde en gedomesticeerde zwijnen. In plaats daarvan weerspiegelen de meeste CNVRs eerder de biogeografie en de demografische geschiedenis van de populaties, zoals dat ook voor andere vormen van variatie (SNPs, microsatellieten) al eerder werd beschreven. De studie zoals in hoofdstuk 2 beschreven betreft een gedetailleerde analyse van CNVRs in wilde en gedomesticeerde zwijnen die waardevolle nieuwe inzichten verschaft in de evolutionaire dynamiek van CNVs tijdens domesticatie en aanpassing aan veranderende omstandigheden.

Verscheidene studies hebben aangetoond dat SVs, vooral die waarbij genen gedupliceerd of verloren raken, het belangrijkste mechanisme is in gen- en genoomevolutie. Met name studies aan primaten hebben inzicht verschaft in de rol van CNVRs in de evolutie van soorten. In het algemeen echter wordt ons inzicht in de rol van CNVR in soortvorming beperkt doordat er niet veel informatie voorhanden is van zeer nauw verwante soorten, m.n. voor gevallen waar het evolutionaire proces van soortvorming feitelijk nog gaande is. Ook hier vormen

zwijnen een uitstekend modelsysteem, zoals is te lezen in hoofdstuk 3. Naast Sus scrofa van het Euraziatisch vasteland komen op de grote eilanden en eilandengroepen in Zuidoost Azië (Java, Borneo, Sulawesi, en de Filipijnen) verschillende andere zwijnensoorten voor die behoren tot het genus Sus. Deze soorten zijn weliswaar morfologisch goed gedefinieerd maar tegelijkertijd in staat onderling nakomelingen te krijgen, wat suggereert dat het hier een nog altijd voortschrijdend proces van soortvorming betreft. Van vijf soorten binnen het geslacht Sus werden de CNVs in kaart gebracht. In totaal werden 1.408 CNVRs gevonden waarbij m.n. CNVRs die overlappen met genen betrokken bij het reukvermogen (Olfactory Receptors, of OR) oververtegenwoordigd zijn. Een vergelijking van een fylogenie van de vijf soorten op basis van de met ORs geassocieerde CNVRs en een fylogenie op basis van CNVRs die overlappen met andere genen laat een interessant verschil zien: waar de fylogenie van de OR-CNVRs de geaccepteerde stamboom van het geslacht Sus weerspiegelt, laat de fylogenie van de niet-OR CNVRs veel meer tekenen van drift en/of vermenging van de verschillende soorten zien. De verklaring hiervoor kan zijn dat de ORs een belangrijke rol hebben gespeeld in de aanpassing aan de lokale omgeving, zoals het geval zal zijn geweest bij het eerste uit elkaar gaan van de verschillende evolutionaire lijnen binnen het geslacht Sus in het Plioceen als gevolg van (tijdelijke) isolatie op verschillende eilanden. Vervolgens, tijdens zeespiegeldalingen in het Pleistoceen en het daarbij behorende verbreken van geografische isolatie, kunnen de genomische gebieden met ORs een – imperfecte - barrière hebben gevormd tegen hybridisatie van de verschillende soorten. De ORs zouden onder dat scenario een directe rol kunnen hebben gespeeld in de soortvorming.

Hoofdstukken 2 en 3 betrof CNVRs die groter zijn dan 5 Kbp, en daardoor gaven deze studies geen inzicht in de kleinere structurele variaties in het genoom. De volgende, natuurlijke, stap was daarom om een alomvattende CNV-studie uit te voeren op gen-niveau. Aangezien de ORs oververtegenwoordigd waren in de

eerdere CNV-studies (hoofdstukken 2 en 3) werd besloten om de studie in hoofdstuk 4 met name op die complexe groep van genen te verrichten. De focus op deze specifieke gen familie staat een accuratere voorspelling van structurele variatie per gen in het OR repertoire toe. Middels DNA-sequentie data van 36 wilde en gedomesticeerde zwijnen uit Azië en Europa werd het belang van selectie en drift in de evolutie van de OR genen bestudeerd. Rond de 60% van de OR-genen blijkt te variëren in het aantal kopieën. De meerderheid van de OR genen die in aantal kopieën variëren vormen clusters van genen in het genoom. Dit sterkt het vermoeden dat zogenaamde 'Non-Allelic Homologous Recombination (NAHR)' een rol speelt in het genereren van variatie in het aantal kopieën. De hoge mate waarin de aantallen kopieën van functionele ORs verschillen, zowel tussen als binnen populaties, suggereert dat deze diversiteit wordt veroorzaakt door selectie. Het was enigszins verrassend te constateren dat de verdeling van het relatieve aantal kopieën van niet-functionele OR genen (pseudogenen) significant afwijkt van de normale verdeling die verwacht word bij neutrale evolutie, dus evolutie in de afwezigheid van selectie. Omdat zowel de functionele als niet functionele OR-genen voor komen in clusters in het genoom is het te verwachten dat NAHR tegelijkertijd de variatie in het aantal kopieën in beide groepen genen heeft veranderd. Dit duidt erop dat naast selectie ook de mate van clustering van gen families in het genoom een belangrijke rol speelt in het genereren van variatie in het aantal kopieën.

In hoofdstuk 5 werd genoomsequentiedata van 42 wilde en gedomesticeerde zwijnen, die vier verschillende populaties uit Europa en Azië vertegenwoordigen, gebruikt om een gedetailleerde kaart te maken van korte inserties en deleties (INDELs) en SNPs. Een half miljoen INDELs (≤10 basen in lengte) werd gevonden per populatie, evenals meer dan zes miljoen SNPs. De SNPs en INDELs zijn verdeeld over het gehele genoom, maar voor beide typen geldt dat meer variatie wordt gevonden aan het begin en het eind van chromosomen ten opzichte van het

midden van chromosomen. Zogenaamde 'Polymerase Slippage' lijkt het belangrijkste mechanisme waarmee INDELs worden gevormd in het varkensgenoom. Gemiddeld werden meer dan 165.000 INDELs in geannoteerde varkensgenen gevonden per populatie, waarbij gemiddeld 422 frame-shift mutaties voorkwamen. Analyse van gen-expressie data liet echter zien dat de meerderheid van die mutaties buiten het werkelijke gen vielen. Het betrof hier dus in meerderheid vals-positieven als gevolg van fouten in de annotatie van genen in het varkensgenoom. Echte frame-shift veroorzakende INDELs zijn dus betrekkelijk zeldzaam, wat te verwachten is omdat hier overwegend tegen zal worden geselecteerd. Het is daarmee aan te raden om de annotatie te controleren alvorens conclusies te trekken met betrekking tot de rol van INDELs bij ziekte en andere kenmerken.

In hoofdstuk 6, tenslotte, worden de belangrijkste bevindingen van dit proefschrift bediscussieerd, met name m.b.t. het belang voor ons begrip over domesticatie en soortvorming. Verder bediscussieer ik ook de sterke en zwakke punten van de huidige methoden voor het detecteren van SVs met behulp van NGS data, en bediscussieer ik de te verwachten trends in de toekomst.

Acknowledgements

I am using this opportunity to express my gratitude to everyone who contributed to the successful completion of my four years wonderful journey of PhD life in this beautiful and vibrant *city of life* sciences. I feel very lucky to be part of WIAS, especially the research team and the project.

I remember the first skype interviews with Martien and Ole, and Richard and Hendrik-Jan followed by personal interviews in Wageningen. First of all, I would like to thank all of you for believing in me and providing me with this great opportunity to be a part of your team. Martien, it was a great learning experience with you. I am sincerely grateful for your aspiring guidance, invaluably constructive criticism, and friendly advice during my PhD.

Ole, you have been a tremendous supervisor for me. I would like to thank you for your support in shaping my knowledge in genomics and for helping me to grow as a research scientist. Without your support, this thesis and I wouldn't be this far. There are no words that can express my gratitude for your guidance, advices, and suggestions during the course of this PhD. Hendrik-Jan you are a great supervisor. Thank you for sharing your experiences and knowledge during meetings, which helped to shape my projects. All of you supported me enormously to improve this thesis.

Thank you to my project members Laurent and Mirte, you are not only my project members but also really good friends. I will always remember our time in old Zodiac, new Zodiac, and Radix. Thanks both of you for helping me to understand genomics in the beginning. Laurent, thanks for your help for my projects, your comments on papers were always very helpful. Thank you also for introducing different bars in Utrecht.

Thank you all the current and past genomics group members, Juanma, Martin, Luqman, Rea, Kyle, Andre, Anna, David, Barsha, Tristan, Marta, Rocky, Petra, Jente, Suvi et al. for your comments during meetings. Juanma, thanks for sharing your experiences and all your help during these four years. I thank you for helping me not to get stressed during the late PhD stage. I miss the coffee breaks. Thanks, a lot for Spanish dinners, you really cook better than any Spanish restaurant!!! Gus, thank you for your help in R and of course introducing different beers and Australian culture. I really enjoyed the Australian day celebration with you and the winter BBQs. Thanks Rea for sharing your experiences and the Greek dishes you cooked for us. Wish you all the best for your future.

Big G, the one and the only Argentine, ops sorry, Uruguayan friend of mine. Thanks for your visits, sharing your lifetime experiences, and “Mate time”. Bert, you are great. I always enjoyed talking to you. Thanks for your support for my project and giving me opportunities to cross the river for birthday parties and BBQs.
Thanks to the Brazilian community, Marcos, Andre and Sandrine. I really enjoyed my time talking to you. Thanks for Brazilian dinners and drinks.
Thanks Hamed, for the Iranian dinners, the yearly ABGC BBQ, and discussions about life.

Thank you all the quantitative genomics friends. I have learnt a lot from all of you. Without you, the PhD life in Wageningen wouldn’t be as lively as it was. Thanks to the lunch group Kyle, Juanma, Jovana, Maulik, Tessa, Claudia, Rocky, Panya, Hooiling, Merina, Rajesh, Suvi, Marta, and Mandy. It was very nice to have you guys around. You guys rock. Chinese dinner nights were always superb, thanks to Mr. Peng, Mandy, Rocky, Lihong, and Tishan.
Jovana and Tessa, thanks for your smile and discussions during coffee time. I always enjoyed talking to you guys. Nancy, Marzieh and Mahlet, it was always nice to tease you guys. Thanks for the African dinners. The `Danish` guys Mahmoud, Hadi and Setegn, you hard working fellows thanks for sharing your experiences.

Thanks to Ada and Lisette for your assistance. You made my PhD life a lot easier.

Thanks to Tycho and Nastia, very good friends of mine out of my office, for organizing trips and enjoying dinners together during the weekends. I really appreciated the time with you in Wageningen.

Thank you my Nepalese friends in Wageningen. Arun, Deepak, Ekaraj, Ram, Rajesh, Shailendra. Prerana, Mary et al., for making me at home during festivals.

Thanks Enrico, my manager at Roche, and Dheeraj for arranging a very smooth transition between my PhD and the Postdoc.

Special thanks to my mom and dad for their everlasting love and care. Words cannot express how grateful I am to you for all of the sacrifices that you have made on my behalf. Your prayer for me was what kept me going on. Thanks to my dear sisters Sapana, Srijana, and Bhawana, brother Amrit, and brother-in-laws Binod, Yuvaraj and Lars, who are also my best friends, for their support, love, and care throughout these years. Thanks for always being there. Thanks to my dearest

nephew Aaryavarta and nieces Ayushree and Maya for your sweet words and smiles, which made my visits unforgettable. At the end, I would like to express my appreciation to my love Marina who was always my support. Your unconditional love and sacrifices are what sustained me this far.

Curriculum Vitae

About the author

Yogesh Paudel was born on 18th May 1984 in Bara district of Nepal. He finished his bachelor degree (B.Sc.) in computer science from St. Xavier's College, Tribhuvan University, Nepal in 2006. After working as an IT officer in an international organization for 5 months, he went to Sweden to pursue his master degree (M.Sc.). He completed his M.Sc. in bioinformatics from Stockholm University in June 2009. During his masters, he received the Japanese Science and Technology Academy award and went to the National Institute of Biomedical Innovation (NIBIO), Osaka, Japan to investigate "one-dimensional structural features and their relationships with conformational flexibility in helical membrane proteins". Soon after finishing his master degree, he received the Stockholm University research grant and worked for 3 months on projects related to protein structure prediction at the Center for Biomembrane Research (CBR), Stockholm University, Sweden. After finishing his 3 months stay at CBR, he moved to Germany and worked as a research assistant in the Institute for Biostatistics and Informatics in Medicine and Ageing Research (IBIMA) at the University of Rostock. In October 2010, Yogesh enrolled as a PhD student at Wageningen University and moved to the Netherlands. He submitted his PhD thesis in October 2014. During his PhD, he investigated different types of structural variations in pig genomes, published in this thesis. After his PhD, his interest in human diseases led him to take a position as a postdoctoral researcher in F. Hoffmann-La Roche AG, Basel, Switzerland, where he is a part of a biomarker team investigating the genetics of schizophrenia and related neurological disorders.

List of publications
Publication under review or in preparation

Y Paudel, O Madsen, H J Megens, R P M A Crooijmans, M A M Groenen. Assessing the functional impact of short insertions and deletions in pig genomes. (Submitted)

Y Paudel, O Madsen, H J Megens, L A F Frantz, M Bosse, R P M A Crooijmans, M A M Groenen. Analysis of copy number variation of olfactory receptors (ORs) in pig (Sus scrofa) genomes suggest a role for genetic drift and selection in the evolution of ORs. (Submitted)

Y Paudel, O Madsen, H J Megens, L A F Frantz, M Bosse, R P M A Crooijmans, M A M Groenen. Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. *BMC genomics* (Under review)

W Y Leung, T Marschall, **Y Paudel**, L Falquet, H Mei, A Schoenhuth, T Y Maoz. The new face of SV detection and tool development for all species SV-AUTOPILOT: Structural Variation AUTOmated PIpeLine Optimization Tool. *BMC genomics* (Under review)

Peer-reviewed publication

M Bosse, H J Megens, L A F Frantz, O Madsen, G Larson, **Y Paudel**, N Duijvesteijn, B Harlizius, Y Hagemeijer, R P M A Crooijmans, M A M Groenen. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature communications* (2014). 10.1038/ncomms5392

L A F Frantz, J G Schraiber, O Madsen, H J Megens, M Bosse, **Y Paudel**, G Semiadi, E Meijaard, N Li, R P M A Crooijmans, A L Archibald, M Slatkin, L B Schook, G Larson, M A M Groenen Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome Biology* (2013). 10.1186/gb-2013-14-9-r107

Y Paudel, O Madsen, H J Megens, L A F Frantz, M Bosse, J W M Bastiaansen, R P M A Crooijmans, M A M Groenen Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC genomics* (2013). 10.1186/1471-2164-14-449

A E Codina, **Y Paudel**, L Ferretti, E Raineri, H J Megens, L Silió, M C Rodríguez, M A M Groenen, S E Ramos-Onsins, M P Enciso. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. *BMC Genomics* (2013). 10.1186/1471-2164-14-148

M Bosse, H J Megens, O Madsen, **Y Paudel**, L A F Frantz, L B Schook, R P M A Crooijmans, M A M Groenen. Regions of homozygosity in the porcine genome: Consequence of demography and the recombination landscape. *PLOS Genetics* (2012). 10.1371/journal.pgen.1003100

The Swine Genome Consortium. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* (2012). 10.1038/nature11622

M J Pfeiffer, M Siatkowski, **Y Paudel**, S T Balbach, N Baeumer, N Crosetto, H CA Drexler, G Fuellen, M Boiani. Complexity of the proteome of metaphase II mouse oocytes: towards a comprehensive definition of the 'reprogramome'. *Journal of Proteome Research* (2011). 10.1021/pr100706k

A Som, C Harder, B Greber, M Siatkowski, **Y Paudel**, G Warsow, C Cap, H Schöler, G Fuellen. The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. *PLOS ONE* (2010). 10.1371/journal.pone.0015165

S Ahmad, Y Singh, **Y Paudel**, T Mori, Y Sugita, K Mizuguchi. Integrated prediction of one-dimensional structure features and their relationships with conformational flexibility in helical membrane proteins. *BMC Bioinformatics* (2010). 10.1186/1471-2105-11-533

Training and education

Training and education

The Basic Package (3.0 ECTS)	**Year**
WIAS Introduction Course	2010
Ethics and Philosophy in Life Sciences	2011

Scientific Exposure (14.0 ECTS)	
International conferences	
WIAS science day, Wageningen, The Netherlands	2011
WIAS science day, Wageningen, The Netherlands	2012
Society of Molecular Biology and Evolution, Dublin, Ireland	2012
The Netherlands Bioinformatics Conference, The Netherlands	2013
NGS methods for identification of mutations and large structural variants, Lausanne, Switzerland	2014
Seminars and workshops	
CBSG Technology Symposium "Advances in life-science Technologies", Wageningen, The Netherlands	2010
Genomics and Animal Breeding, Wageningen, Wageningen	2011
WIAS seminar: from DNA to daily practice	2011
NBIC: BioAssist Programmers Meeting, Utrecht, The Netherlands	2013
NGS methods for identification of mutations and large structural variations	2014
Presentations	
Otto Warburg International Summer School and Research Symposium, Berlin, Germany (Poster)	2011
WIAS science day, Wageningen, The Netherlands (Poster)	2012
Society of Molecular Biology and Evolution, Dublin, Ireland (Poster)	2012
The Netherlands Bioinformatics Conference, The Netherlands (Poster)	2013
NBIC: BioAssist Programmers Meeting, Utrecht, The Netherlands (Oral)	2013
NGS methods for identification of mutations and large structural variants, Lausanne, Switzerland (Oral)	2014

In-Depth Studies (9.0 ECTS)	
Disciplinary and interdisciplinary courses	
MPI Otto Warburg International Summer School and Research Symposium on Evolutionary Genomics - Berlin Germany	2011
SIB winter school: In the intersection of bioinformatics and medicine - Kandersteg, Switzerland	2013
EMBL Advanced Course: Next Generation Sequencing Data Analysis - Heidelberg, Germany	2013
CSHL: Computational & Comparative Genomics - New York, USA	2013
PhD students' discussion groups	
PhD – paper discussion group	2011-2012
Professional Skills Support Courses (4.0 ECTS)	
Scientific Writing	2012
Communication course	2013
Techniques for Writing and Presenting Scientific Papers	2014
Research Skills Training (6.0 ECTS)	
Preparing own PhD research proposal	2011
Didactic Skills Training (14.0 ECTS)	
Supervising practicals and excursions	
Genomics WUR	2011-2013
Supervisor theses	
Animal Breeding and Genomics 4- MSc students	2011-2013
Management Skills Training (1.0 ECTS)	
Organization of seminars and courses	
Organization of NGS course and workshop, Lausanne, Switzerland	2014
Education and Training Total	**50 ECTS**

Colophon

This work was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC Grant agreement no 249894 (SelSweep project).

The cover of this thesis was designed by Yogesh Paudel.

Printed by GVO drukkers en vorngevers B.V./Ponsen & Looijen, Ede, The Netherlands.