Am I repeating myself? Determining the repetitive landscape of the pig X chromosome

Sarah Quigley

A thesis submitted for the degree of MSD Biological Sciences

Department of Life Sciences

University of Essex

Date of submission 10/2021

Abstract

Mammalian sex chromosomes evolved from a homologous pair of autosomes via the acquisition of a major sex determining gene. This event led to the suppression of recombination between the chromosomes and consequently their independent evolution. Both X and Y have been observed to accumulate amplified gene families in mice and humans, but it remains unclear to what extent this is true in other mammalian species. The pig X chromosome was recently sequenced to a high quality and in this study, it was investigated to determine the presence and extent of potential amplicons. LASTZ was used to align the pig X sequence to itself and identify regions of similarity that could represent ampliconic sequences. Further analysis revealed many of the similarities to be interspersed along the chromosome, and in some cases particularly clustered around the centromere. This distribution of hits suggests many of the similarities to be the result of known repetitive elements which commonly cluster within centromeric and pericentromeric regions. There were also some regions showing segmental duplications and gene duplications. Further investigation of the hits through NCBI BLAST revealed them to be fragments of LINE-1 (Long Interspersed Nuclear Element 1) retrotransposons, orthologues of duplicated genes, and uncharacterised loci. Mammalian X chromosomes are often enriched with LINE-1s which are thought to play a role in X-inactivation and controlling gene expression. Some of the duplicated genes are of potential interest for further analysis as their function is unknown. At this stage, no independent amplicons have been found in the pig X chromosome, in contrast to what has been observed in humans and mice. However, the current assembly contains gaps between contigs and unmasked repeats which may obscure where potential amplicons might be found. The data from this study can be used to improve the current annotation of the pig X chromosome.

Acknowledgements

I would like to thank my supervisor Dr Ben Skinner and lab group PhD member Ellie Watson for helping with my project by providing academic and emotional support. I would like to thank Dr Antonio Marco for his help and advice during our board meetings. I would also like to thank my father, Danny Quigley, for providing emotional and financial support allowing me to dedicate all my time and efforts to producing this project.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

I. How is sex determined? The classification of sex determination systems

Where do babies come from? We all asked this question, and then we would ask why there are boys and girls? Often, we achieve a 'satisfactory' explanation in school when learning about sperm and egg cells and that's that. There is much more to it than that. Sex in this context is the term used to broadly explain the biological determination of male or female offspring. Sex determination mechanisms are genetic, environmental, reversible, and irreversible (Kent et al., 1996). Working individually or in combination they produce male or female offspring. Biosynthesis of sex steroids in some fish uses a combination of genetic, hormonal and environmental factors (Devlin & Nagahama, 2002). Reptile sex determination occurs during the critical period of gonad development influenced by temperature (Janzen & Paukstis, 1991). The Jacky dragon (A.*muricatus*) for example, requires extreme temperatures to promote the fitness of females and intermediate temperatures promote male fitness (Warner & Shine, 2008). Where the sexual phenotype and sexual genotype are mismatched (i.e. genetic components for female offspring yet phenotypic traits expressed for male offspring) there can be sex reversal in vertebrates, known as hermaphrodites, and this sex reversal can be triggered through environmental and hormonal treatments (Baroiller & D'Cotta, 2016).

Genetic sex determination is a broad term encompassing; two factor systems, multiple factor systems, Haplodiploidy systems, and cytoplasmic sex determination (Charlesworth, 2002). Contrary to popular belief not all genetic sex determination is controlled via the nucleus. Some genes can be found in the cytoplasm, often inherited from the mother, creating a selective pressure favouring female offspring aka cytoplasmic sex determination (Werren, 1987). Examples of animals which may experience these selection pressures are; mites, fruit flies, butterflies, and some plants (Werren, 1987). Haplodiploidy systems as suggested by Hamilton (1964) refers to the fertilisation or lack of fertilisation of a female egg in sex determination, most often observed in

eusocial species such as honeybees. Haplodiploidy sex determination produces haploid males from unfertilised eggs and diploid females from fertilised eggs (Rautiala et al., 2019).

Two factor and multiple factor systems involve meiotic mechanisms:- a diploid gamete is divided into four haploid gametes for the exchange of genetic material via genetic recombination (Bell, 1982). These sex determination systems have been reported as the XY system (multiple factor) and the ZW system (two factor). Both systems have a homogametic and heterogametic sex. In XY systems the male is heterogametic (XY) and females homogametic (XX); interestingly, contrasted by ZW systems with a heterogametic female (ZW) and homogametic male (ZZ) (Irwin, 2018). Both systems have a diminished chromosome (Y/W) via chromosome degeneration through reduced recombination (See 1.0 below) (Ezaz et al., 2006). The ZW system has been observed in species such as chickens and birds where production of male offspring is dosage sensitive, meaning the Z-linked genes require two doses for male development (Smith et al., 2009). Male development in XY systems however, is stimulated by the sex determining region on the Y (*SRY*) (Berta et al., 1990). There are a number of similarities between the XY and ZW systems – such as similar heterogenous/homogenous sex determination and a degraded sex determining chromosome (Y/W) suggesting they evolved from a common ancestor. Their evolution from this common ancestor however appears to have occurred independently as suggested by the lack of significant homology between XY and ZW (Fridolfsson et al., 1998).

II. Evolution of mammalian sex chromosome systems

From the late 1800's the growing field of research into mammalian sex chromosomes began (Abbott et al., 2017) providing an insight into their evolution from a homologous pair of autosomes ~180million years ago (Cortez et al., 2014; Skinner et al., 2016). This divergent evolution increased their susceptibility to unique evolutionary forces such as mutation rates, sexual selection, recombination rates, and genetic conflict influencing sex chromosome structure and function (Bachtrog, 2013; Bachtrog et al., 2011; Ellis et al., 2011).

Prior to sex chromosome evolution, the offspring sex was determined through a sex determining pathway involving the gene family *SOX*. The gene *SOX9* has been seen to play a role in testes development where *SOX3* acts as a repressor of *SOX9* during sex determination in females (Schwarz, 2000). A mutation, translocation, or duplication of genes involved in this sex determining pathway led to the acquisition of a major sex determining function onto an autosome, beginning sex chromosome evolution (Beukeboom & Perrin, 2014). Figure 1.1 taken and altered from Graves (2006); shows the acquisition of the sex determining region on the Y (*SRY,* yellow in Figure 1.1&B), evolved from the autosomal *SOX3*, leading to the first stage of sex chromosome evolution (Berta et al., 1990; Graves, 1998; Herpin & Schartl, 2015). *SRY* and *SOX3* share a homologous DNA binding domain - HMG (high mobility group) box; suggesting an evolutionary relationship (Foster et al., 1994). SRY and SOX3 are alleles found at the same locus forming proto-sex chromosomes; the *SRY*-bearing autosome becoming the proto-Y chromosome and the *SOX3*-bearing autosome becoming the proto-X chromosome, as shown in Figure 1.1A&B (Vallender & Lahn, 2004).

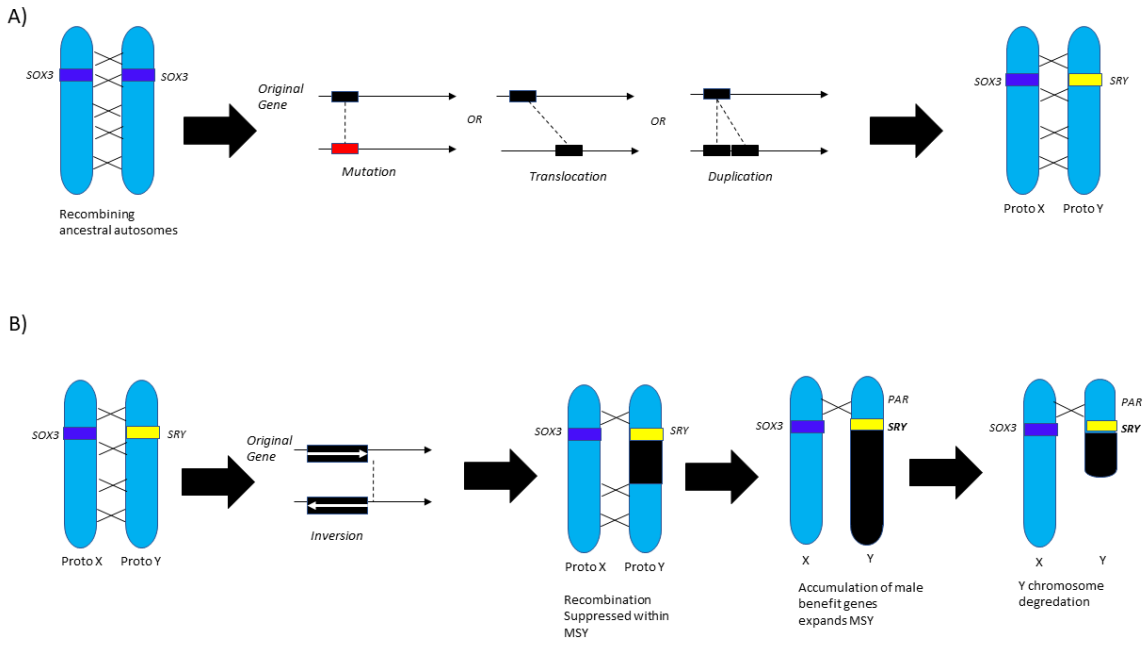

Figure 1.1 Evolution of recombining ancestral autosomes (recombination shown with crosses) to the conserved X chromosome and degraded Y chromosome._A) The autosomal gene SOX3 (dark blue) experienced either a mutation event (represented as the black gene changing to red), a translocation event (represented as the black gene being moved along the arrow), or a duplication

event (represented as two of the black gene showing a tandem repeat). These events can lead to the evolution of SOX3 to SRY, the sex determining region on the Y (yellow). The acquisition of SRY forms the proto sex chromosomes. B) Once proto sex chromosomes, a large-scale inversion can occur on the Y. This inversion leads to reduced recombination within the MSY – male specific region (in black) containing SRY (in yellow). Over time the accumulation of male benefit genes leads to an increase in the size of the MSY. Recombination does not occur within the MSY but continues within the Pseudoautosomal region labelled PAR. Finally, the reduced recombination prevents removal of problematic mutations consequently resulting in gene loss within the Y chromosome.

Antagonism between functional properties of new mutations led to polymorphisms becoming established, these polymorphisms (Figure 1.1B) were favoured by natural selection as they resulted in recombination arrest forming the non-recombining regions on the Y (NRY, Figure 1.1B in black) also known as the male specific region (MSY) (Abbott et al., 2017; Beukeboom & Perrin, 2014;Rice, 1987). Where recombination continues at the sex determining locus, the evolution of closely linked sexually antagonistic mutations can both be beneficial to one sex and detrimental to the other (Rice, 1987). These mutations, in the presence of recombination, spread to increase their transmission to the preferred sex, regardless of the collateral (Ma et al., 2018). Sex chromosome antagonism established polymorphisms, such as genetic modifiers and large-scale inversions (often on the Y chromosome) to reduce recombination. These polymorphisms generated closer linkage of male beneficial mutations to the sex determining locus; preventing spreading of the mutations and thereby reducing their influence on female population fitness (Charlesworth, 2017). Reduced recombination occurs within the NRY and this region increases in size with the accumulation of male benefit genes (Graves, 2006). The larger the NRY the more recombination is suppressed between the chromosomes as represented in Figure 1.1B. Muller (1918) first noticed the reduction in recombination between the X and Y chromosomes although noted this observation was not concordant in the XX chromosome pairs.

Reduced recombination results in the independent evolution of the X and Y chromosomes as seen where the ancestral gene content differs greatly with coverage around 3% in the Y and 98% in the X (Disteche, 2016). Sex chromosome inheritance mechanisms also differ between the X and Y chromosomes. Homozygous females inherit one X from each parent whereas

heterozygous males inherit the X from their mother and a clonal copy of the Y from their father (Hughes et al., 2005; Ross et al., 2005).

Genetic erosion in the Y chromosome (see 1.0) causes a loss of homology between the X&Y chromosomes. A region of homology however remains. Genes having survived degradation, or translocations of autosomal segments onto the end of the sex chromosomes, form the pseudoautosomal region (PAR); where recombination can occur outside of the NRY (Gil-Fernández et al., 2020). Figure 1.1B shows homologous recombination still occurs outside the NRY within the blue areas on the Y chromosome representing the PAR. This is required between the X and Y for ensuring proper pairing and segregation of the chromosomes during meiosis preventing male infertility and XY aneuploidy (Terje Raudsepp & Chowdhary, 2015). During meiosis, the chiasmata is also essential for the correct alignment and separation of the chromosomes, the formation of these chiasmata is facilitated by the repair of double strand breaks (DSB) during prophase (Murakami & Keeney, 2008). Research into DSB frequency found DSBs were infrequent in the autosomes of mice yet they were dense within the sex chromosome PAR promoting homologous recombination (Acquaviva et al., 2020).

After recombination suppression, evolutionary distance on the X chromosome can be measured through evolutionary strata (Sandstedt & Tucker, 2004). The existence of six evolutionary strata shows no less than six events occurred in the genetic divergence of the sex chromosomes (Disteche, 2016; Skaletsky et al., 2003). Stratum one begins the evolutionary story with the initation of recombination suppression and stratum 6 representing the newest chapter. Any genes within the same stratum began differentiating into X and Y homologs at about the same time (Lahn & Page, 1999). It would then be likely that XY genes seen in different strata would have independently evolved. *SRY* is found in the oldest stratum (one) serving as further conformation of the theory that the acquisition of the sex determining region triggered differentiation of X and Y chromosomes (Bachtrog, 2013). Reduced recombination fortified this

differentiation thus forming morphological and genetic differences creating the X and Y chromosomes we know (Bachtrog, 2006).

III. Y chromosome independent evolution led to degredation and gene loss

Recombination arrest triggered the independent evolution of the Y chromosome and thereby its degradation and genetic erosion (Charlesworth & Charlesworth, 2000). Y chromosome gene loss yielded a size disparity where it became 3 times smaller than the X (Charlesworth, 1996; Vallender & Lahn, 2004). Oxidative stress in the testis and a lack of repair enzymes may have contributed to Y degradation, also where sperm require more divisions than eggs providing further opportunity for chromosome damage (Graves, 2006).

The Y chromosome structure is comprised of the PAR, the short arm (Yp) and the long arm (Yq) separated by the centromere on the acrocentric chromosome (Colaco & Modi, 2018; Raudsepp et al., 2012). The acquisition of *SRY* onto the Y chromosome led to the formation of the male specific region of the Y (MSY, Figure 1.1). Mammalian MSYs contain large repeat units (often arranged in tandem repeats) known as amplicons along with X-degenerate (X-d) sequences (Hughes et al., 2010). The X-d regions contain single copy homologues of X-linked genes that do not recombine yet they serve as evidence for the once homologous XY chromosomes and their autosomal ancestry (Hughes et al., 2010).

Y-borne genes present their own story of the stages of Y chromosome degeneration. Genes early in their erosion are still partially activated unlike the more ‘mature’ degradation witnessed in Y pseudogenes with active X homologues. The finale of degradation results in a complete loss of some X homologue genes on the Y chromosome (Graves, 2002). Chronicles of this phenomenon often surmise recombination arrest is the cause of degeneration, yet why would this be the case? Muller (1918) suggested recessive mutations could not be expressed due male heterozygosity; preventing selection against the recessive allele therefore where deleterious mutations occur there

is decay of Y-linked genes (Saxena, 2000). Fisher (1935) disputed this theory when assessing gene mutation rates and equilibrium; it was suggested Muller's theory would not make such an impact as the deleterious mutant gene would be likely to be selected against. Homologous X-linked loss of function genes were also suggested by Fisher to prevent accumulation of deleterious genes prior to Y degradation (Rice, 1996).

Muller (1964) later devised a theory for the degradation of the Y chromosome which came to be known as Muller's ratchet. This theory stated that where deleterious mutations occur, in the absence of recombination, there will be substantial accumulations of the mutant alleles (Gabriel et al., 1993). Consequently these accumulations decrease the maximum potential fitness for the population and in some conditions this could continue to reduce indefinitely (Butcher, 1995). Muller's ratchet would then be associated with a decline in genetic diversity and linked to the degeneration of the Y chromosome due to the loss of genes through deleterious mutations (Gordo et al., 2002).

Working in conjunction with Muller's ratchet is the Hill-Robertson interference (Hill & Robertson, 1966). The Hill-Robertson (HR) interference suggests selection acting on one site reduces the effectiveness of selection on a linked site (Comeron, 2005). Recombination reduces the effect of interference therefore in the context of sex chromosomes with reduced recombination the HR effect is increased (Roze & Barton, 2006). HR interference refers to beneficial or deleterious mutations within the same haplotype affecting each other's selection, therefore in accordance with Muller's ratchet any mutations occurring in a population will accumulate and become fixed (da Silva & Galbraith, 2017).

However, Muller's ratchet was brought into question as the rate of these accumulations appeared to be implausible unless occurring within an incredibly small population (Charlesworth, 1978; Charlesworth, 1996). If in some case Muller's ratchet did prove to be true in a population;

Charlesworth (1978) believed it would only apply where the mean number of deleterious mutations had increased but did not necessarily lead to a fixation of the mutation. Clonal inheritance of the Y chromosome from father to son provides a small population size; therefore deleterious mutations in the population, with reduced recombination, results in accumulations of the mutant alleles (Gordo & Charlesworth, 2000). This accumulation leads to a decline of fitness and reduced recombination prevents the removal of these 'problematic' mutations (Hughes & Page, 2015).

Degradation of the Y chromosome is associated with genetic hitchhiking. This involved fixations of deleterious alleles where there are favourable Y-linked mutants (Charlesworth, 1991). In other words; positive directional selection of the favourable allele linked with a deleterious locus, will fix both the allele and locus in the presence of reduced recombination (Bachtrog, 2013).

IV. X chromosome dosage compensation mechanisms balance Y degradation

Mammalian X chromosomes are composed of the X-conserved region (XCR) and X-added region (XAR) (Mácha et al., 2012). In 1967, Ohno (2013) proposed mammalian X chromosomes to be highly conserved due to homology shared between X-linked and autosomal mendelian genes. This homology was not seen in the Y chromosome as a result of genetic erosion. A comparative genomics study by Mueller et al (2013) found homology between single copy genes in humans and mice, showing conservation, however ampliconic genes shared little homology (see section 1.0I)

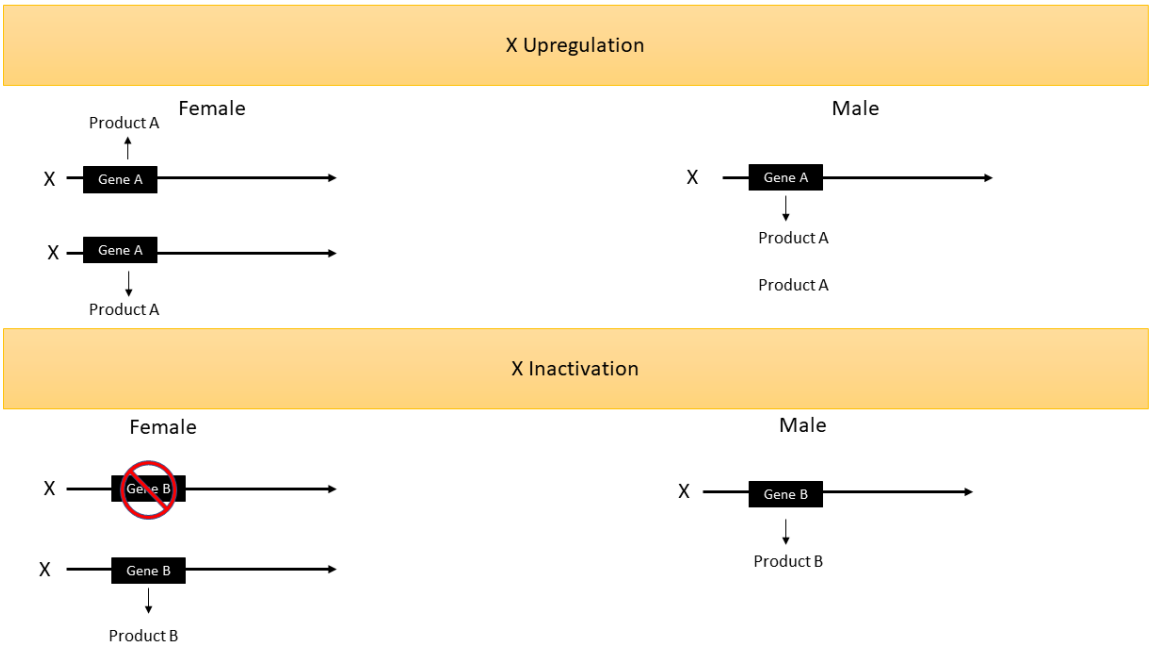

Figure 1.2 Visual representation of dosage compensation in the X chromosome. The dosage of X-linked genes in males is half that of females due to gene loss in the Y chromosome, consequently, this causes the gene product to be half that of females. To compensate for this, dosage compensation mechanisms have been developed. One mechanism is upregulation of the X in males to meet the gene content of that in females. This is seen where the male gene A produces twice the amount of product from the single gene resulting in both males and females having two units of product A. Alternatively there is X inactivation. This mechanism inactivates genes on one of the female X chromosomes. In this case, gene B from the male produces one product B unit and only one of the female's genes produces a product therefore each have one unit of product B.

Haploinsufficiency is prominent where genetic erosion in the Y leaves males with a single X-linked copy of a gene (Berletch et al., 2010; Sahakyan et al., 2017). Dosage compensation aims to restore the balance of gene expression between males and females. The mechanisms can either involve upregulating the single copy of the gene (males) or inactivating one X chromosome (females): represented in Figure 1.2 as adapted from Charlesworth (1996) (Deng et al., 2013; Gupta et al., 2006). X upregulation requires DNA sequences or epigenetic mechanisms to control transcription (Deng et al., 2013). In some cases, X gene upregulation can be influenced by autosomal gene dosage suggested where upregulation occurs upon fertilisation (Nguyen & Disteche, 2006). The X to autosome ratio then influences X gene expression where upregulation can balance the ratio below 1 (Deng et al., 2009). The justification of dosage compensation in mammals is debated; some studies suggested X-upregulation was a response to sexual dimorphism in gene expression and one study suggested X inactivation evolved not to reduce female

overexpression but as an imprinting mechanism to silence growth inhibiting genes in embryos (Disteche, 2016).

X chromosome inactivation (XCI) involves condensing the chromosome to heterochromatic foci leaving it transcriptionally static, aka the "Barr Body" (Jenuwein & David Allis, 2001). The selection of the chromosome to silence can either be random (most common) or predetermined via imprinting (Heard et al., 2004). XCI has three main stages; initiation, propagation of inactivation throughout the X chromosome, and maintenance of inactivation independent of the X inactivation centre (Xic) (Clemson et al., 1996). The initiation stage often appears to involve the X inactive specific transcript (*Xist*) gene as shown where *Xist* is activated just prior to XCI (Brown et al., 1991; Panning & Jaenisch, 1996). Supporting the suggestion of *Xist*'s involvement in XCI is it's localisation to the Xic which controls the active state of the chromosome; as blocking the Xic is required for the chromosome to remain active (Brown et al., 1991; Rastan & Robertson, 1985). *Xist* RNAs recruitment initiates transcriptional silencing and allows for protein recruitment (Dechaud et al., 2019). Gene expression is altered by histone modifications, the recruited proteins, and protein complexes for the regulation of DNA replication and chromosome segregation (Zhang & Reinberg, 2001).

Propagating this initiation across the X chromosome requires Xist RNA to be spread along the chromosome facilitated by "way stations", consisting of transposable elements known as LINEs (long interspersed nuclear elements) (Chow et al., 2010). Way stations containing LINEs enhance *Xist* RNA attachment to the chromosome allowing for heterochromatization (Matsuno et al., 2019). The study by Simon et al., (2013) supported the correlation found between *Xist* spreading and LINE density where *Xist* localisation was found to coincide with Xic looping contacts. LINE enrichment on the human X chromosome was around twice the average for the genome supporting the suggestion that LINEs are involved in XCI (Popova et al., 2006). Further confirmation of the involvement of LINEs in *Xist* spreading was the detection of LINEs within the

proximity of *Xist* silenced genes (Sousa et al., 2019). The first stage of *Xist* spreading was the contact with gene rich regions then moving towards the more distal gene-poor inter-regions (Simon et al., 2013). LINE-1 is overrepresented on the X chromosome although there have been suggestions that the evolutionary older LINEs were involved in the assembly of the *Xist*-dependent repressive domain followed by the newer LINES spreading *Xist* (Garcia-Perez et al., 2016).

XCI does not cover the entire chromosome; some genes escape inactivation, as grouped: within the PAR, dosage-sensitive X-linked genes with Y-linked homologues, variable escape genes, and facultative escape genes (Carrel & Brown, 2017; Lyon, 1962; Ma et al., 2018). Escape genes remain expressed in both the active and inactive X chromosome alleles, although the X inactive (Xi) genes are lower in their expression (Berletch et al., 2011). This lowered expression of escapee genes was also observed by Chen et al (2016) reporting expression decreased with XCI progression although it was still detectable. Expression of escaped genes is higher in females than males leading to sexual dimorphism (Fang et al., 2019). The presence of abnormal escape genes can lead to deleterious phenotypes including infertility, intellectual disability, immune diseases, and cancer (Fang et al., 2019). Contrary to the implication, escaping XCI is not rare with 15% of human genes escaping inactivation and this escape also results in some bi-allelic expression (Berletch et al., 2015). Further evidence of these active escape genes is the depletion of *Xist* spreading in these genes although there is some *Xist* enrichment in the surrounding regions of these escapees (Simon et al., 2013). Genes expressed in the Xi lack the epigenetic markers that represent inactivated genes as concordant with their active status (Fang et al., 2019).

The location of these escape genes is so that they are distanced from repressive genomic elements (Fang et al., 2019). The distancing from repressive genomic elements removes these genes from *Xist* RNA and LINE elements therefore the inactivation of the genes is slower allowing for their escape (Sousa et al., 2019). Escape genes were found to be located in the X short arm in strata 3 and 4 which were at that time the newest strata (Disteche, 1999). Some XCI escape genes

showed to be tissue specific and were depleted in epigenetic repressors such as the histone modifications involved in XCI but enriched in marks associated with active transcription (Disteche & Berletch, 2015). These tissue specific genes have higher expression in females as seen by Tukiainen et al (2017) exhibiting higher sex-biased expression than other Xi genes although this bias is not restricted towards females as some PAR escape genes were male biased. Disteche (1999) suggested dosage may not be important for why some genes escape XCI or the dosage is important for a sex-specific function where escapees may need a higher expression in females for ovarian function.

Alternatively, where XCI occurs as intended, a consequence may be haploinsufficiency- the loss of one allele (possibly through deletion or loss of function -inactivation) leaving the remaining and active allele to be insufficient (Nguyen et al., 2019). Haploinsufficiency is the reason dosage compensation mechanisms are required where Y degradation (section 1.0) leads to insufficient allele dosage (Figure 1.2) (Yazdi et al., 2020). The balancing of gene expression through XCI in females can lead to haploinsufficiency where there is a single X chromosome and insufficient alleles for their products (Skuse, 2005). Escape genes are often essential for normal development and in some cases the escape genes can be haploinsufficient (Chaligné & Heard, 2014). X chromosome haploinsufficiency is associated with chromosomal abnormalities and diseases such as Turner syndrome (TS) (Gibson et al., 2018). TS in humans involves haploinsufficiency of genes within the PAR1 (humans have two PARs – 1&2) region where genes are encoded for receptors such as GTP binding proteins, ATP transporter, and transcription factors (Bakalov et al., 2009). These encoded receptors allow for the TS transcription factors to be involved in the abnormal insulin secretory response characteristic of TS (Bakalov et al., 2009). In 2008, the human genome had around 300 haploinsufficient genes identified where 60% were located between segment duplications (Dang et al., 2008).

V. Structure and composition of the pig sex chromosomes

The genome of the domestic pig (*Sus scrofa*) is organised into 18 autosomal pairs and the X and Y chromosomes; 38 chromosomes in total (Gustavsson, 1988). The karyotype arrangement consists of twelve autosomes with two arms, six acrocentric chromosomes, a submetacentric X, and a submetacentric Y (Adega et al., 2005). The initial *Sus scrofa* genome assembly was performed by Groenen et al., (2012) using bacterial artificial chromosome (BAC) and whole-genome shotgun (WGS) sequences and this assembly included the X chromosome although not the Y chromosome. Groenen et al., (2012) discovered 95 novel repeat families within the pig reference genome arranged in 5 long interspersed nuclear elements (LINEs), 6 short interspersed nuclear elements (SINEs), 8 satellites and 76 long terminal repeats (LTRs).

Olfactory receptors are one of the largest mammalian gene families (Nguyen et al., 2012). Studies of the mouse and human have shown the olfactory receptor genes to be located in all chromosomes except chromosome 18 in humans, chromosome 12 in mice, and the Y chromosome in both species (Olender et al., 2008; X. Zhang & Firestein, 2002). This suggests the presence of the olfactory receptor within the X chromosome with no Y homologue in mammals. The function of the olfactory receptor has been suggested to be involved in signalling, detecting and releasing pheromones in the female pig (Baum, 2012). Nguyen et al., (2012) annotated the pig reference genome discovering the lowest percentage of olfactory receptor pseudogenes compared to any studied species. The pig has one of the largest active olfactory receptor gene family repertoires and this gene family is the largest in the pig genome suggesting it has been highly amplified (Paudel et al., 2015).

Selection pressures particularly affect the pig X chromosome especially at the end of the short arm where the PAR is located (Ma et al., 2014). These selection pressures (natural and artificial selection) shaped the relative fitness of the domestic pig (Zhang et al., 2020). The PAR also provides a location for XY homology along with a region towards the q-terminus of the Xq

(Skinner et al., 2013). Further investigation of selection in the pig X chromosome showed evidence of strong selective sweeps in the population (Ai et al., 2015). Pigs are an important vector for studies into domestication and artificial selection; their ancestor- the European wild boar highlights their predominant lineage within European domestic swine (Larson et al., 2007). The selective sweep in the X chromosome may have occurred pre-domestication as seen where there are large regions of homology between the domestic pig and its ancestor the wild boar (Rubin et al., 2012).

To this point the pig reference genome comprised of all 18 chromosomes and the X chromosome. Further assembly and gene annotation of the pig X chromosome was performed by Skinner et al., (2016) including a first draft of the pig Y chromosome structure and gene content. Prior to this assembly the composition of the Y (having large quantities of constitutive heterochromatin) was determined by studies assessing C-band patterns (chromosome staining, highlighting constitutive heterochromatin associated with repeats) suggesting the presence of ampliconic regions (Adega et al., 2005). Skinner et al (2016) confirmed the presence of repeats on the Y chromosome suggested by the C-band maps and mapped the majority of its single copy genes on the short arm and the repetitive regions on the Y long arm. Repetitive regions have not only been observed in the pig Y chromosome, the presence of potential amplicons have also been mapped onto the pig X chromosome although to a lesser extent (Iacolina et al., 2016; Skinner et al., 2013). Warr et al., (2020) worked on updating the pig genome assembly and suggested there is still more complexity within the genome yet to be annotated, for example, within the repetitive regions.

The MSY region of mammalian Y chromosomes is often generalised to be gene poor, although the sequenced MSY in pigs shows to be one of the few to counter this generalisation by being gene dense (Janečka et al., 2018). Multicopy genes contribute to this density including the gene families in the pig are the *HSFY* gene family (as discussed in 1.VI) and *TSPY* (the testis-specific protein Y-linked) gene families (Li et al., 2013; Moretti et al., 2016). Skinner et al., (2015)

discovered elements of high copy numbers of genes such as with the *HSFY* gene family with at least two forms of the gene family at over 100 copies. The *TSPY* gene family on the other hand is a multicopy gene family within the pig genome and is arranged in tandem repeats (Vodicka et al., 2007). *TSPY* is associated with a number of cancers and male infertility (in humans) potentially by influencing the ability of the individual to produce sufficient and physiologically competent spermatazoa (Vodicka et al., 2007). The study by Janečka et al., (2018) showed a large number of X-Y multicopy MSY genes shared between species where all were present in the pig genome, these genes included; *SRY*, *ZFY*, *TSPY*, *DDX3Y, UTY*, *USP9Y, BCORY, AP1S2Y, ZRSR2Y, and CUL4BY* (Table 1.1).

Table 1.1 Y linked multicopy genes found in pig MSY region showing their abbreviated and full scientific names

Gene Abbreviation	Gene Name	Reference
SRY	Sex determining region on the Y	(Berta et al., 1990).
ZFY	Zinc finger gene	(Koopman et al., 1991)
DDX3Y	DEAD Box helicase 3	(Ramathal et al., 2015)
USP9Y	Ubiquiin Specific Peptidase 9	(Ramathal et al., 2015)
UTY	Ubiquitously transcribed tetratricopeptide repeat gene	(Nailwal & Chauhan, 2017)
AP1S2Y	P-1 complex sub- unit sigma-2	(Smeds et al., 2019)

The nature of highly repetitive sex chromosomes, such as the pig Y chromosome, leave them predisposed to deletion and duplications consequently affecting testis development and function (Krausz & Casamonti, 2017). The master sex determining gene *SRY* acts as the switch gene for sex determination in pigs as it does in other mammals (Kashimada & Koopman, 2010). Many of the genes found in the pig MSY discussed above such as *DDX3* have both X (*DDX3X*) and Y (*DDX3Y*) homologues and have been shown to be expressed through different stages of spermatogenesis in other mammals (Rauschendorf et al., 2014). These genes found in the pig MSY are important in spermatogenic function and due to their location are transmitted from father to son without X recombination (Teixeira et al., 2019). These MSY genes are involved in sex determination, the maintenance of testicular development, and spermatogenesis. The ancestral Y-

linked genes *UTY*, *USP9Y*, *DDX3Y* and *ZFY* are reported to regulate protein production through each stage of DNA translation into functional products in humans (Bellott et al., 2014).

VI. Genomic conflict can occur between amplified gene families

The divergent evolution of the X and Y chromosomes resulted in reduced recombination (as discussed in 1.II and 1.0) and this gave rise to an increase in sequence amplifications (Lucotte et al., 2018; Skinner et al., 2016). Sequence amplifications are often formed as identical tandem arrays and inverted repeats known as amplicons with around 99% similarity (Bellott & Page, 2009; Skaletsky et al., 2003). An example includes the inverted repeats located on the human Y chromosome (IR1 and IR4); IR1 contains two almost identical amplicons comprising of the azoospermia factor c (AZFc) which is an nearly entirely ampliconic region of the chromosome essential for spermatogenesis (Lange et al., 2013).

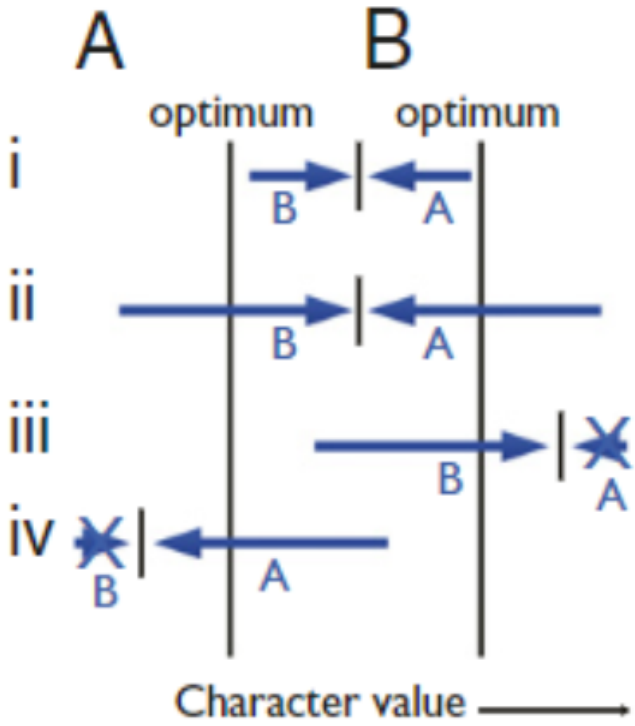

Figure 1.3 Taken from Frank & Crespi (2011) showing genes in conflict. Two genes in conflict, in this case named A and B, will act against one another. i) A and B are acting opposingly with equal force. This solution compromises to a value which is at neither gene's optimum level. ii) The overhanging reaction from i would be the reaction of one gene to increase its force, consequently leading to the opposing gene to increase its own force leading to the solution to stay in the same level. iii and iv) The result of

removing one of the forces through knockout is seen. iii) The opposing force from A has been knocked out and therefore allows for B to reach its optimum unopposed and in many cases this is registered as an overexpression of B. iv) alternately the vice versa is true with B knockout and A is unopposed and can reach its optimum, again, appearing to be overexpressed.

There are 3 main classes of the azoospermia factor named from a to c found to be ampliconic within the human Y chromosome (Repping et al., 2003; Skaletsky et al., 2003). Ye et al., (2018) associated AZF amplicons with reproductive diseases where deletions in the amplicons showed reductions in fertility, although amplicons outside of the AZF regions also showed this relationship. Sequencing of mammalian Y chromosomes aided in the suggestion that the function of ampliconic sequences is to counterbalance the degradation of the MSY and gene loss (Hughes et al., 2012; Soh et al., 2014).

Human and mouse X-linked amplicons were predominantly expressed in the testis implying they play a role in selection for gene expression during or after male meiosis (Nam et al., 2015). One third of independently acquired X-linked genes, in humans and mice, showed to be single-copy or multicopy and two thirds ampliconic (Mueller et al., 2013). Multicopy gene families have been found to be expressed within the testis in a number of mammals such as the mouse, human, horse, bull and pig, and this suggests they have a role in spermatogenesis (Moretti et al., 2020). Testis expressed ampliconic genes within the sex chromosomes are proposed to be involved in mechanisms such as meiotic drive and therefore are involved in genomic conflict (Dutheil et al., 2015).

When the acquisition of a reproductive advantage may also pose detrimental consequences to the fitness of parts of the genome, this is known as genomic conflict. For example, genetic elements within an individual, genes in separate individuals or similar genomic regions such as those in males and females have opposing selection (Rice, 2013). Fitness of a gene as defined in Gardner & Welch (2011) is the number of gene positions in the subsequent generation that receive their genetic material from that gene position in the current generation. Intragenomic conflict involves elements invested in promoting their own transmission rate through interference with the

transmission of other alleles or replicating disproportionately to the rest of the genome (Crespi & Nosil, 2013). Intragenomic conflict can be defined in three categories, origin conflict -where they came from, destination conflict -where they are heading, and situation conflict -their current position (Gardner & Francisco, 2017). Intragenomic conflict in the context of sex chromosomes is the conflict between maternal and paternal origin genes within the offspring (Haig 2006). Gardner & Francisco (2017) classified examples of intragenomic conflict concerning the three categories e.g. imprinted genes are an example of origin conflict, meiotic drive and transposable elements are examples of destination conflict, and sex chromosomal genes are involved in situation conflict.

A beneficial allele that incurs a cost for other alleles may be acted upon by natural selection to find a balance between the advantages and disadvantages; in the germ line there are different values placed upon the fitness of males compared to females therefore there is intragenomic conflict between these sexually antagonistic traits (Hitchcock & Gardner, 2020). These antagonistic traits are represented in Figure 1.3 taken from Frank & Crespi, (2011). In the absence of one of the antagonistic alleles, through mutation or knockouts, leads to the dominating expression or overexpression of the remaining allele (Frank & Crespi, 2011).

An example of intragenomic conflict between the sex chromosomes is the occurrence of sex ratio skews. According to Fisher (1958) the role of natural selection will ensure evolution towards achieving an equilibrium of a 1:1 ratio of male to female offspring referred to as “Fishers principle” – seen in mammals. This was debated by Hamilton (1967) whom agreed this may be true for the homogametic sex although it is not applicable in the heterogametic sex. For example, say a there is a mutation in the Y chromosome causing the Y-bearing sperm to always succeed in fertilisation, then a male bearing this mutation produces nothing but sons (Hamilton, 1967). The nature of the Y chromosome inheritance, where sons inherit a clone of the chromosome from their fathers, would mean this mutation would persist throughout generations. Fisher (1958) argued that males and females contribute in equal proportion to the gene pool and the genes involved in

controlling sexual proportion segregate in a mendelian fashion (Carvalho et al., 1998). The equal contribution is not the case where X-linked genes in males are entirely maternal, therefore the number of male offspring does not affect the fitness of the gene; and if the X-linked gene is involved in controlling sexual proportion there will be a bias towards females (Carvalho et al., 1998). Y chromosome clonal inheritance leaves Y-linked gene fitness to be determined by the number of sons in a population and the inheritance of a gene leading to a higher proportion of males will be able to spread through meiotic drive (West, 2009). Asymmetric transmission of X and Y linked genes leaves the sex chromosomes prone to genomic conflict due to their unequal reproduction in genomic components (Rice, 2013).

These models of the selective forces in fitness maximisation follow the mendelian pattern of inheritance; but, contrary to these there are selfish genetic elements (SGE) or Transmission distorters (TD) involved in genomic conflict (Zanders & Unckless, 2019). One well known transmission distorter is the t-complex distorter (tcd) controlling the level of distortion of the transmission ratios in mice (Silver, 1989). This t-complex distorter is not sex linked as it is found on chromosome 17, yet, it provides a good example of meiotic drivers acting within their own self-interest (Kelemen & Vicoso, 2018). The t-complex is found in all subspecies of the house mouse *Mus musculus* and inherited at frequencies higher than 90% (Carroll et al., 2004). The high abundance of this mutation is due to its ability to trigger four tandem inversions resulting in recombination suppression and promote its own transmission (Gummere et al., 1986; Sugimoto, 2014). This is consistent with the suggestion that distorter genes require linkage with responder genes for survival; and where recombination can occur this link may be broken and the trait cannot be transmitted (Schwander et al., 2014). When expressed, the t mutation presents the phenotypic characteristic of tailless mice (Sugimoto, 2014). Sex chromosomes with defined reduced recombination provide an ideal environment for meiotic drive, in regard to the function of TDs and SGEs.

SGEs act to increase their own transmission into the next generation, particularly where the genes involved are functionally related to sex allocation as seen in mice (Scott & West, 2019). This work has challenged the conceptualisation of the genome as a highly coordinated network with mendelian segregation, to what we now recognise as SGEs that act to promote their own transmission at the expense of other genes in the genome (Agren, 2016). The sex chromosomes are prime examples of SGEs showing some form of drive. The evolution and selection for suppressors of that drive are required to balance the conflict (Burt & Trivers, 2009). An evolutionary arms race is a consequence of SGEs manipulating meiosis to increase their transmission to over 50% of offspring (Ellison & Bachtrog, 2019). The meiotic drivers and their suppressors have distinct genetic signatures and often have high rates of lineage-specific duplications and gene amplifications (Ellison & Bachtrog, 2019). As discussed above these amplifications occurred in the sex chromosomes where there was reduced recombination as a strategy to counteract degradation. However homologous recombination between the amplicons leaves them prone to rearrangements and has been found to result in deletions and spermatogenic failure (Repping et al., 2003).

Few mammals have sufficiently assembled X and Y reference genomes, although, humans and mice have a complete X and MSY assembly (Soh et al., 2014). Studies of the mouse Y chromosome have provided evidence that Y-linked genes are involved in spermiogenesis; including sperm motility and development and carry genetic information essential for fertility (Paria et al., 2011; Touré et al., 2004). As discussed the sequence amplifications counterbalanced the decay of the mammalian MSY particularly in the mouse where the MSY is almost entirely ampliconic (Hughes et al., 2010; Skaletsky et al., 2003). Post-meiotic intragenomic conflict in the sex chromosomes of mice led to the amplification of gene families-*Slx, Sly* and *Ssty* (Cocquet et al., 2009). These gene families are ampliconic gene families. The mapping of the mouse Y

chromosome localised the *Sly* gene families to the long arm (Yq) and the sex determining gene *Sry* to the short arm (Yp) (Soh et al., 2014).

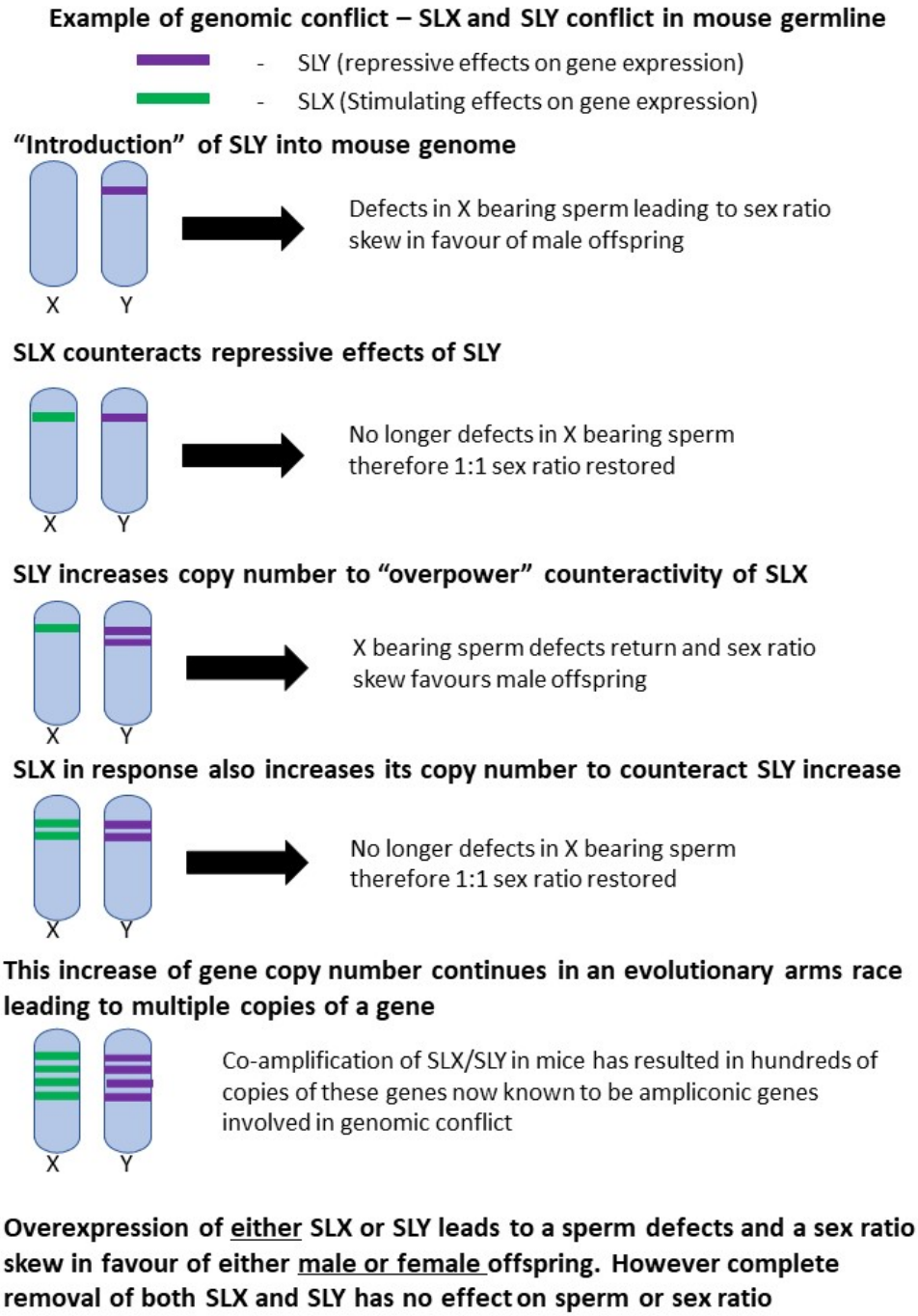

Figure 1.4 An example of Genomic conflict as seen in the mouse lineage where there is conflict between the regulator genes SLX and SLY. The Y-linked SLY represses sex chromosome transcription in spermatids the opposite is also true of the X-linked SLX which stimulates sex chromosome transcription. The presence of both SLY and SLX is required for normal spermatogenesis, although, absence of both genes also ensures normal spermatogenesis therefore the function of these genes is to promote their own transmission. Often a deletion of either SLY or SLX results in sperm abnormalities and therefore a sex ratio skew in favour of one sex.

Studies found X and Y homologues of the ampliconic regulator genes – *Slx* and *Sly* which have been observed to be in conflict where they have opposing functions – see Figure 1.4 (Rathje

et al., 2019). *Sly* is essential for normal spermatogenesis and deletions result in the overexpression of *Slx* genes and a sex ratio skew in favour of females (Bachtrog, 2014; Rathje et al., 2019). *Sly* deletions occur when there is a deletion in the MSYq (Y long arm) leading to sperm head abnormalities and reduced Y bearing sperm motility, although larger deletions can lead to a removal of *Sry* which leads to male infertility (Ward & Burgoyne, 2006). *Sly* acts as a suppressor of gene expression as shown where deletions favour female offspring showing the role of *Sly* as an SGE (Bachtrog, 2014). These coamplified suppressors and drivers (*Sly* and *Slx* respectively) have antagonistic functions therefore they compete directly with each other acting as SGEs (Figure 1.4). Consequently *Slx* knockouts form a bias towards male offspring where X bearing sperm have abnormalities such as abnormal spermatid elongation, reduced sperm counts and reduced motility (Cocquet et al., 2010; Ellison & Bachtrog, 2019). A recent study by Moretti et al., (2020) confirmed the well-studied theory of antagonism between *Sly* and *Slx* determining that the proteins encoded by the ampliconic genes compete for the interaction with the product of the multicopy gene *Ssty*.

Recently a study looked into the ampliconic gene content of the bull MSY; and is significant as it is one of the first studies to suggest a large number of amplifications may be a broad characteristic of mammalian MSYs rather than being a peculiarity in mice (Hughes et al., 2020). In the past, multicopy gene families have been observed in a variety of mammalian MSYs however many Y assemblies are still incomplete. The mouse and bull both having multicopy gene families in their MSY regions were compared and due to the lack of homology between the two (*Sly* and *Ssty* in mouse; *HSFY*, *ZNF280BY*, and *TSPY* pseudogenes in bull) it was determined these amplifications occurred independently (Hughes et al., 2020). The HSFX and HSFY gene family exists in bull, humans and pig; where the bull was found to have 79 copies of *HSFY* with 11 copies of its co amplified X-homologue *HSFX* (Hughes et al., 2020; Skinner et al., 2015). *HSFY* is the heat shock transcription factor Y-linked; in humans this is found within the azoospermia factor b

(AZFb) region (Tessari et al., 2004). Heat shock factor genes act as transcriptional regulators for the functions of the heat shock proteins by interacting with their highly conserved heat shock element genes - although *HSFY* itself does not bind to heat shock element genes (Kichine et al., 2012; Rabindran et al., 1991). Stress stimuli such as high temperatures or oxidative stress can damage proteins and this is where the heat shock proteins step in to repair the damage or in some cases these proteins can induce apoptosis (Pirkkala et al., 2001).

The altered expression of the *HSFY* variant has shown to result in deterioration of spermatogenesis and may subsequently alter cell differentiation in sperm (Sato et al., 2006). Deletions of *HSFY* appear to link to male infertility although not solely responsible; there could be a role in the phenotypic expression of some alleles that influence flagellar formation (Kichine et al., 2012). The coamplified *HSFY* and *HSFX* with high intra-family nucleotide identity may have ongoing gene conversion and the independent evolution of the bovine X and Y suggests there is no ongoing recombination (Hughes et al., 2020). The highly amplified copies of *HSFY* show a potential for higher variable copy numbers of the *HSFY* gene family in mammals compared to the low copies found so far in humans, mice, and cats (Skinner et al., 2015). *HSFY* amplification has been seen in pigs, occuring independently, suggesting the gene may be 'prone' to amplification (Skinner et al., 2015). *HSFY* has testis specific expression in the bull, human and pig suggesting it plays similar role in spermatogenesis across the mammalian lineages (Yue et al., 2014).

VII. Comparing mammalian sequence assemblies and alignment methodologies

Comparative genomic studies rely on high-quality genome assemblies. Large genomes can prove challenging to assemble due to greater financial cost and increased complexity from repetitive content (Koepfli et al., 2015). The assembled genome is assigned a value called the N50 summarizing the assembly quality using the length of the scaffold (or contig) representative of the shortest scaffold (or contig) needed to cover 50% of the genome (Mäkinen et al., 2012). When comparing N50 values to assess the assembly quality, the larger the value the more complete the

genome assembly and where there are fewer scaffolds there is a more complete assembly (Jensen-Seaman et al., 2004). Once the assembly quality is determined, the next stage in genomics analysis studies is often genome annotation and genome alignment; these methods allow for determining the conservation of a genome and genetic relationships (Koepfli et al., 2015).

Many mammalian sex chromosomes have been assembled, although not all of the assemblies include a Y chromosome. This is likely due to the small size and highly repetitive content of the Y chromosome leading it to be difficult to assemble (Liu et al., 2019). To avoid the complexities of sequencing the Y chromosome a female of the species is often studied (Liu et al., 2019).

Scaffold numbers and N50 values were obtained for the mammalian sex chromosome assemblies from the NCBI database (*Genome - Assembly - NCBI*, n.d.) (Table 1.2 and Table 1.3) along with the number of known genes from the Ensembl databank (*Ensembl Genome Browser 101*, n.d.). The number of sex chromosome scaffolds compared to the number of unplaced scaffolds in the genome can also give an insight into assembly quality along with the N50 value. Understanding the current assembly qualities for mammalian sex chromosomes provides an insight into the work that has previously been done, the research which still needs to be performed, and which species will provide accurate comparisons for other studies. Mammalian sex chromosome assemblies were compared using the number of scaffolds and their N50 values (Table 1.2 and Table 1.3). Fewer scaffolds often suggest a more complete assembly however many unplaced scaffolds in the genome and a smaller scaffold number suggests an incomplete assembly. This was represented as a fraction where the denominator of the fraction represents the unplaced scaffolds for the entire genome and the numerator represents the scaffolds placed to the chromosome (Table 1.2 and Table 1.3).

The assemblies for the mouse and human sex chromosomes appear to be of high quality and are thought to be nearly complete assemblies. This is shown where they have a high scaffold

N50 for their sex chromosomes, low scaffold numbers, and few unplaced scaffolds reflecting the high-quality assembly of both sex chromosomes associated with the human and mouse genome (Webster et al., 2019). Comparing mammalian sex chromosome assemblies show the pig sex chromosome still requires improvements. However, the updated 11.1 assembly improved upon the deficiencies of the 10.2 assembly, this is reflected with the introduction of the Y chromosome and the high scaffold N50 in the updated X assembly which is now comparable to the mouse and human complete assemblies (Warr et al., 2020).

Table 1.2 Mammalian X and Y chromosome assembly qualities. The number of scaffolds for the X and Y chromosome is represented using a fraction where the numerator represents the number of scaffolds placed to the X or Y chromosome and the denominator represents the total number of unplaced scaffolds in the entire genome. The X and Y gene column show the total number of known genes in the chromosome, including the coding genes, non-coding genes, and pseudogenes.

Mammal	Number of X scaffolds	Number of Y scaffolds	X scaffold N50	Y scaffold N50	X genes	Y genes
Chimpanzee **Pan_tro_3.0**	203/42786	24/42786	25963054	2357686	1387	153
Human **GRCh38**	9/126	14/126	34966268	6276129	2408	566
Macaque **Mmul_10 & 5**	8/2625	3/2625	59398400	7945854	1506	67
Vervet-AGM **chlSab2**	123/1432	26/1432	130038232	6181219	1122	18
Mouse **GRCm38.p6**	15/22	20/22	28873115	17801351	2619	1560
Rat **Rnor_6.0 & 3.1**	66/578	29/578	11448667	5447879	1492	48
Blue Whale **mBalMus1.v2**	1/106	1/106	129330052	2427530	952	24
Domestic Yak-**LU_Bosgru_v3.0**	1/383	1/383	136336377	26357969	1061	197
Cattle **ARS-UCD1.2, Bos indicus 1.0 and Btau 5.0.1**	1/31	1/31	88516652	39421065	1363	192
Pig **Sscrofa11.1**	3/583	79/583	65063120	12044924	1311	129

Table 1.3 Mammalian X chromosome assembly qualities. The mammalian species in this table only have X chromosome assemblies and do not currently have an assembled Y chromosome. The number of scaffolds for the X chromosome is represented using a fraction where the numerator represents the number of scaffolds placed to the X chromosome and the denominator represents the total number of unplaced scaffolds in the entire genome. The X and Y gene column show the total number of known genes in the chromosome, including the coding genes, non-coding genes, and pseudogenes

Mammal	Number of X scaffolds	X scaffold N50	X genes
Bonobo Panpan1.1	66/134720106	4911963	1296
Crab eating macaque Macaca_fascicularis_5.0	14/7107	93731725	1295
Gibbon Nleu_3.0	234/15567	106279408	1154
Gorrilla gorGor4	14/39767	94278388	1315
Mouse lemur Mmur_3.0	219/2913	144923918	1266
Olive Baboon Panu_3.0	667/63213	405323	1252
Algerian mouse SPRET_EiJ_v1	2/5382	148322235	1868
Northern American deer mouse HU_Pman_2.1	1/8499	134369076	1113
Prairie vole MicOch1.0	8/6303	11399579	440
Rabbit OryCun2.0	5/3218	33683086	979
Shrew mouse PAHARI_EIJ_v1.1	2/2552	148379591	1547
Arabian camel CamDro2	1/23402	119869850	971
Canada lynx mLynCan4_v1.p	1/6445586	122081366	886
Cat Felis_catus_9.0	28/4142	78035653	1212
Cow ARS-UCD1.2	1/2180	139009144	1132
Dog CanFam3.1	2/3228	123773608	1364
Horse EquCab3.0	1/4668	128206784	1157
Lion PanLeo1.0	1/8041	81249997	444
Sheep Oar_rambouillet_v1.0	1/2613	153341986	1243
Yarkand deer CEY_v1	1/17893	134243713	1151

Following the assembly comes the alignment stage where the target and query sequences are compared to detect any homologous regions. These alignments can either be between the same DNA sequence (self-alignment) or different sequences (Hickey et al., 2013). Multiple genome comparison and alignment tools are used in the identification of biological similarities and

differences in genomes; the information gathered from these analyses provides insight into phylogeny, functional regions, and gene predictions (Treangen & Messeguer, 2006). There are two main forms of alignment algorithms: global and local as seen in Table 1.4. Global alignments of whole genomes are used when comparing two very closely related species (Chain et al., 2003). Complications can arise in global alignments where the size of genomes can be larger than millions of nucleotides, although this can be overcome using some alignment programs that assess maximal matches which are combined into local alignment chains (Delcher et al., 1999). Local alignments are a straightforward comparison between two sequences, making fewer assumptions than global alignment (i.e. the species are closely related), and also allows for finding all pairs of genes or evolutionarily constrained elements between two genomes, transposons and other repeats, and any other similarities (Batzoglou, 2005). Another type of alignment mentioned in Table 1.4; hierarchal whole genome alignment splitting the whole genome alignment into a set of multiple global alignments (Dewey, 2019).

Performing a local alignment can use the basic local alignment tool (BLAST); a fast program using short read algorithms which estimates the number of matches that would have occurred (the expect value) (Madden, 2003). Short reads often do not carry a large amount of information therefore require to be pieced together to reconstruct the biologically important information (Canzar & Salzberg, 2015). Single read mapping complicates the mapping of repetitive regions; as a single read can map to multiple repeats in the genome, leading to the true position of a read being ambiguous and ultimately forming inaccurate or incomplete alignments (Lee & Schatz, 2012). Overcoming this uses a tool to mask repeat elements prior to the alignment using RepeatMasker (Smit et al., n.d.) which screens DNA for transposable elements, satellites and low-complexity DNA sequences (Tempel, 2012). The masking of the repeats involves replacing them with either Ns, Xs, or lowercase letters (Tarailo-Graovac & Chen, 2004).

Two main alignment types were identified in Table 1.4; Pairwise and Multiple. Both Pairwise and Multiple alignments identify regions of similarity that indicate functional, structural or evolutionary relationships (*Pairwise Sequence Alignment Tools < EMBL-EBI*, n.d.). The main contrast between the two is that Pairwise alignments compare two sequences whereas Multiple alignments compare three or more sequences. Pairwise alignments have three known methodologies; dynamic programming, dot-matrix, and 'word' methods (Mount, 2003). Table 1.4 shows the most common methodology in the selected alignment strategies is dynamic programming which can be used in both pairwise alignments and multiple sequence alignments.

Homology assessments using the dynamic programming algorithm aligns the sequences to identify matches, mismatches, and indels to calculate an optimal score providing a mathematical solution (Eddy, 2004). Although, dynamic programming is a computationally demanding methodology (Eddy, 2004). This mathematical algorithm discards poorly conserved initial and terminal fragments although it cannot exclude differing internal fragments (Arslan et al., 2004). This results in an alignment formed of well-conserved fragments artificially connected by poorly conserved or even unrelated fragments. This is problematic as it may be biologically inadequate, although, this can be overcome through modifying the scoring matrix (Arslan et al., 2004). Dynamic programming can be applied to multiple alignments although this has been limited to 3-4 sequences due to the computational demands for memory becoming too large (Lipman et al., 1989). Although according to Langmead & Salzberg, (2012) these computational demands are solved when using modern processors with greater processing power providing fast, accurate, and memory-efficient gapped sequence alignments.

Table 1.4 Comparison of different chromosome alignment methodologies.

METHOD	ALIGNMENT TYPE	PAIRWISE OR MULTIPLE	READ LENGTH	METHOD	REFERENCE
BLAST	local	Pairwise	Short	Dynamic Programming	(Madden, 2003)
M-GCAT	Hierarchal WGA	Multiple	Long	Dynamic Programming	(Treangen & Messeguer, 2006)
STELLAR	Local	Pairwise	Long	Dynamic Programming	(Kehr et al., 2011)
DIALIGN	Global	Multiple	Long	Motif finding	(Brudno et al., 2003)
BURROWS-WHEELER	local	Multiple	Short	Dynamic Programming	(H. Li & Durbin, 2009)
BOWTIE2	local	Multiple	Short	Dynamic programming	(Langmead & Salzberg, 2012)
BLASTZ	local	Pairwise	Long	Dot matrix	(Schwartz et al., 2003)
LASTZ	local	Pairwise	Long	Dot matrix	(R. S. Harris, 2007)

Pairwise alignments provide a more simplistic method for visualising similarities between two sequences through a graphical representation known as the dot-matrix plot method (Sonnhammer & Durbin, 1996). A graph is produced with one sequence on the X axis and the other sequence on the Y axis and a dot marked wherever there are similarities (Shu & Shan Ouw, 2004). Identification of an alignment in the dot plot is seen where there is a diagonal of continuous dots, this diagonal may be broken in cases where mutations have occurred, or shifted where indels are present (Lee & Peng, 2018). Graphic presentations such as the dot-matrix are ideal for analysing long sequences of genomic DNA especially when containing repetitive regions (Huang & Zhang, 2004).

The ampliconic content of the pig X chromosome will be measured in this study therefore a self-alignment will be performed to detect areas of homology. These alignments will only use two sequences at a time therefore pairwise alignments are sufficient (BLASTZ, LASTZ, STELLAR, and BLAST). Finding ampliconic regions requires longer reads as used in BLASTZ,

STELLAR, and LASTZ (Table 1.4). Finally, as discussed, the dot-matrix output is ideal for visualising repetitive regions in the genome (BLASTZ or LASTZ). Due to the nature of LASTZ being an improved pairwise alignment method that is the successor to BLASTZ, LASTZ is preferable. LASTZ provides a wider range of seeding choices, reduced memory requirements (allowing larger sequences to be compared), and multiple output formats (Harris, 2007).

The successor to BLASTZ is tuned to be more sensitive although this comes with the drawback of being a slower aligner (Armstrong et al., 2019). LASTZ has several stages within its process; a raw alignment (the immediate product which may contain overlapping matches in the aligned sequences) and net alignment (the chaining of matches that discards all but the highest scoring chains with a single match); where the net alignment identifies homologous elements in the raw alignment (Gao & Miller, 2014). Repeat-masked sequences are excluded by the LASTZ software during the seeding stage of the alignment which is ideal for our study (see methods) (Harris, 2010). LASTZ allows for the comparison of an entire sequence such as the pig X chromosome against another full sequence such as the pig X /Y or another mammalian X (Huang et al., 2012).

VIII. Wider implications of determining ampliconic genes in the pig X chromosome

The domestic pig (*Sus Scrofa*), a member of the Suidae family, was one of the first domesticated animals providing insight for the effects of domestication; and evolutionary studies with their ancestor (wild boar) alive in the present day (Chen et al., 2007). The pig family Suidae evolved from other *Sus* species around 4 million years ago spreading from southeast Asia across most of Eurasia (Paudel et al., 2013). The domestication of the Suidae family in particular provides a platform for studies into the domestication of modern mammalian lineages where 3 divergent clusters of mitochondrial sequences suggested independent domestication events in the lineages (Larson et al., 2005). Phenotypic variations in the modern pig were a consequence of domestication leading to differentiation from the pig ancestor through variations in morphological characteristics

to reproduction and behaviour (Groenen, 2016). This shows the pig to be a good model for understanding the consequences of selection and genetics particularly in the identification of genes and phenotypic traits (Megens & Groenen, 2012).

The characterisation and mapping of the pig genome began in the late 1980s and early 1990s (Rothschild & Ruvinsky, 2011). In 2003, the swine genome sequencing consortium (SGSC) was founded with the aim of advancing biomedical research through obtaining crucial information through the complete sequencing of the pig genome (Schook et al., 2005). There is a great deal of similarity between the pig and human karyotype particularly with regard to its size and composition which means it provides a good model for studies relating to meiotic recombination or genetic disease (Mary et al., 2014). The use of pigs in biomedical research is ideal due to their anatomy, genetics and physiology being closely representative of humans, not only their similarity makes them ideal, but they are also easy to breed where they produce large litters (Bendixen et al., 2010). Some of the current biomedical advances using pigs involve pig to human Xenotransplantation, neurodegenerative disease studies, and production of pharmaceutical treatments for disease (Whyte & Prather, 2011).

The domestic pig holds great agricultural importance as a source of nutrition for humans providing animal based proteins (Chen et al., 2007). The production of suitable meat, disease resistance, and the efficient use of feed are all agricultural considerations that would be benefited by the production of genetically modified pigs (Whyte & Prather, 2011). Genetic modifications in pigs is now a possible and precise practice where genes can be added, modified, or removed which lead to improved health of pigs and improved models for research (Prather et al., 2008).

A study by Gullett et al., (1993) determined the levels of testosterone affected the quality of meat suggesting male pigs produce lower quality meat providing agricultural and economic implications. This phenomenon is known as boar taint. In some cases to prevent boar taint male

piglets intended for pork production are castrated, yet, this poses ethical concerns and other practical solutions are in demand (Lundstörm et al., 2009). Some countries within the European union have now banned castration of male piglets subject to satisfactory solutions to the negative effect posed by testosterone production in uncastrated male pigs; although around 75% of pigs in the EU as of 2019 are still surgically castrated (Bonneau & Weiler, 2019). As of yet, there is no entirely suitable solution for boar taint particularly with ethical constraints of current practices such as castration and culling etc.

Analysis of the pig sex chromosomes will reveal more insights into the evolution of pigs and mammalian lineages, provide more information for biomedical studies regarding human sex chromosomes and allow for advances in the agricultural industry. The presence of the *HSFX*/Y coamplified gene family in both the bull (Hughes et al., 2020) and the pig (Skinner et al., 2015) provides a basis for evolution research. The problem in the agricultural industry regarding testosterone affecting meat quality may be solved if the observed sex ratio skew in mice, caused by the genomic conflict between the amplified gene families *Slx* and *Sly* (Rathje et al., 2019), could be replicated within the pig where gene amplifications have been previously observed on the sex chromosomes (Skinner et al., 2016).

Ampliconic gene families often arise on the sex chromosomes and they are known to be involved in genomic conflict and are candidate genes for sex ratio skews and transmission ratio distortion, especially in mice. The aim of this study is to identify ampliconic genes on the pig X chromosome using existing sequence data and finding potential Y homologues. The objectives of this study are therefore as follows:

I. Align the pig X chromosome to itself to identify regions of similarity
II. Determine regions on the chromosome which may harbour ampliconic genes
III. Characterise genes from potentially ampliconic regions in the X alignments

IV. Analyse and determine possible Y homologues of X-ampliconic genes in the existing data of the Y chromosome of pigs

V. Compare the X gene content of pigs to that of other mammals

2. Methods

I. *Sus scrofa* DNA sequence

Ampliconic sequences are identified as regions of high similarity in the DNA sequence, therefore, the first requirement in this analysis was to obtain the *Sus Scrofa* (pig) DNA sequence. As mentioned above (section I.VII), the current reference genome assembly for the pig has been updated from the Sscrofa10.2 to Sscrofa11.1 (Warr et al., 2020). For the purposes of this study, the pig X DNA sequence was taken from the Ensembl nucleotide database (https://www.Ensembl.org/Sus_scrofa/Info/Index). Bioinformatics analyses such as LASTZ (see 2.IV) require the use of this text-based format- FASTA (Zhang, n.d.).

This analysis requires the removal of known repetitive elements which are not of interest and can obscure any potential novel repeats; Ensembl has the DNA sequences with the repeats removed through the tool RepeatMasker (Smit et al., n.d.). RepeatMasker finds known repetitive elements such as low-complexity DNA sequences and interspersed repeats in the nucleotide bases; the repeats are masked either using normal repeat-masking' where the repetitive elements are masked using N's and soft-masking where the repetitive elements are masked by converting the uppercase bases to lowercases (https://www.Ensembl.org/Sus_scrofa/Info/Index) (Tarailo-Graovac & Chen, 2004).

Deciding whether to employ 'normal' or 'soft' repeat-masking required thinking ahead to later stages of the analysis, considering the LASTZ alignment software (see 2.III and IV). According to the LASTZ documentation, both masking forms are excluded during the seeding stage of the alignment, however they are then treated differently henceforth (Harris, 2010). Lowercase masking is excluded only during the seeding stage of the analysis (see section 2.IV) whereas masking using N's is further penalised in a similar fashion to transversion mismatches. Transversion mismatches are excluded depending on a given extension rule (see 2.IV). The

inclusion or exclusion of the masked regions in later stages of the analysis can hold a substantial impact on the high scoring segment pairs being formed and whether they are kept or removed from the alignment (Harris, 2010).

The search for ampliconic genes in the pig X chromosome is comparable to finding a needle in a haystack; the needle being the amplicon and the haystack being the vast landscape of repetitive elements. Investigating the ampliconic content of the chromosome involved assessing the repetitive content and comparisons between the masked and unmasked data, isolating unannotated repeats, and identifying their nature. All things being considered, the N masking appeared to be preferable to lowercase masking allowing for known repeats to be completely removed and true comparisons to be made between the masked and unmasked data.

II. Preliminary assessment of inverted repeats in pig X chromosome using IUPACPAL

As mentioned previously, sequence amplifications and ampliconic genes are often identified as tandem arrays and inverted repeats (IRs) (Bellott & Page, 2009; Skaletsky et al., 2003). The human genome has had many inverted repeats identified and some of the largest and most homologous IRs have been detected in the X chromosome (Warburton et al., 2004). Considering these factors, creating an index of the inverted repeats proved useful in observing the repetitive content of the pig X chromosome.

The IUPACPAL alignment software was employed, working similarly to EMBOSS palindrome although detecting inverted repeats overlooked with the EMBOSS tool (Alamro et al., 2021). IUPACPAL significantly reduces the runtime compared to EMBOSS according to Alamro et al., (2021). Once installed, the IUPACPAL tool can be used as a command line function (installation procedure is followed on the Github documentation https://github.com/steven31415/IUPACpal). IUPACPAL requires given parameters including

minimum and maximum length of the palindrome, maximum gap between palindromes and the number of mismatches allowed. The command line syntax used can be found here: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/IUPAC_analysis.

For the purposes of this analysis, the minimum palindrome length was set at 50bp, it is known that palindromes which are involved in inverted duplications and amplifications often occur through cruciform cleavages, and hairpin formations (Darmon et al., 2010). Short palindromes, under 50 bp, often do not form these structures and more often serve as binding sites such as TATA boxes therefore are not of interest in this study (Ganapathiraju et al., 2020). The maximum length however was set at only 1000bp, even though X chromosome palindromes can be found up to 140kb in length (Jackson et al., 2020; Warburton et al., 2004). This was due to larger values increasing the runtime exponentially, therefore any 'medium' length palindromes of interest could be identified. Analysis of the human genome palindrome content by Ganapathiraju et al., (2020) suggests there were many palindromes between 100-200bp and above 200bp on the human X and there may be palindromes of interest of this length in the pig X. The gap length and maximum mismatches were retained at the default values (100 and 0 respectively); as these values were unknown and increasing them also increased runtime exponentially.

Upon running IUPACPAL on the pig X chromosome sequence, a text-based results format was returned, which was edited to form a table of start and stop positions for further computational analysis. This table was then passed through the same computational analysis stages as the LASTZ alignment data; scripts seem here: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/LASTZ_analysis. The stages from section 2IV onwards were performed on both the IUPACPAL and LASTZ data tables however for continuity the methods focus on the LASTZ example.

III. LASTZ preliminary analysis

As mentioned, the IUPACPAL software could not identify inverted repeats larger than 1000bp within 100bp distance. This was unlikely to identify large palindromes which can sometimes harbour ampliconic genes (Jackson et al., 2020; Warburton et al., 2004). Therefore, we also used LASTZ to produce a fast and low sensitivity overview of any large similarities in the chromosome which may be the result of inverted repeats and palindromes.

This involved using the LASTZ command (see 2.IV), the given parameters as provided in the scripts here: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/low.sensitivity.LASTZ. To clearly visualise where larger alignments may be present; sections of the chromosome were assessed individually covering about 1Mbp regions. The following parameters were used to reduce the alignment sensitivity and runtime: --notransition --nogapped --step=100. The --chain parameter was also used to provide a clear visualisation of 'long' regions of alignments by chaining close high scoring segment pairs into a high scoring path.

The output of the LASTZ alignment produced a dot plot showing any chained alignments and their orientation. If any regions were of interest, they were noted and considered, these regions were investigated throughout the LASTZ analysis process as follows:

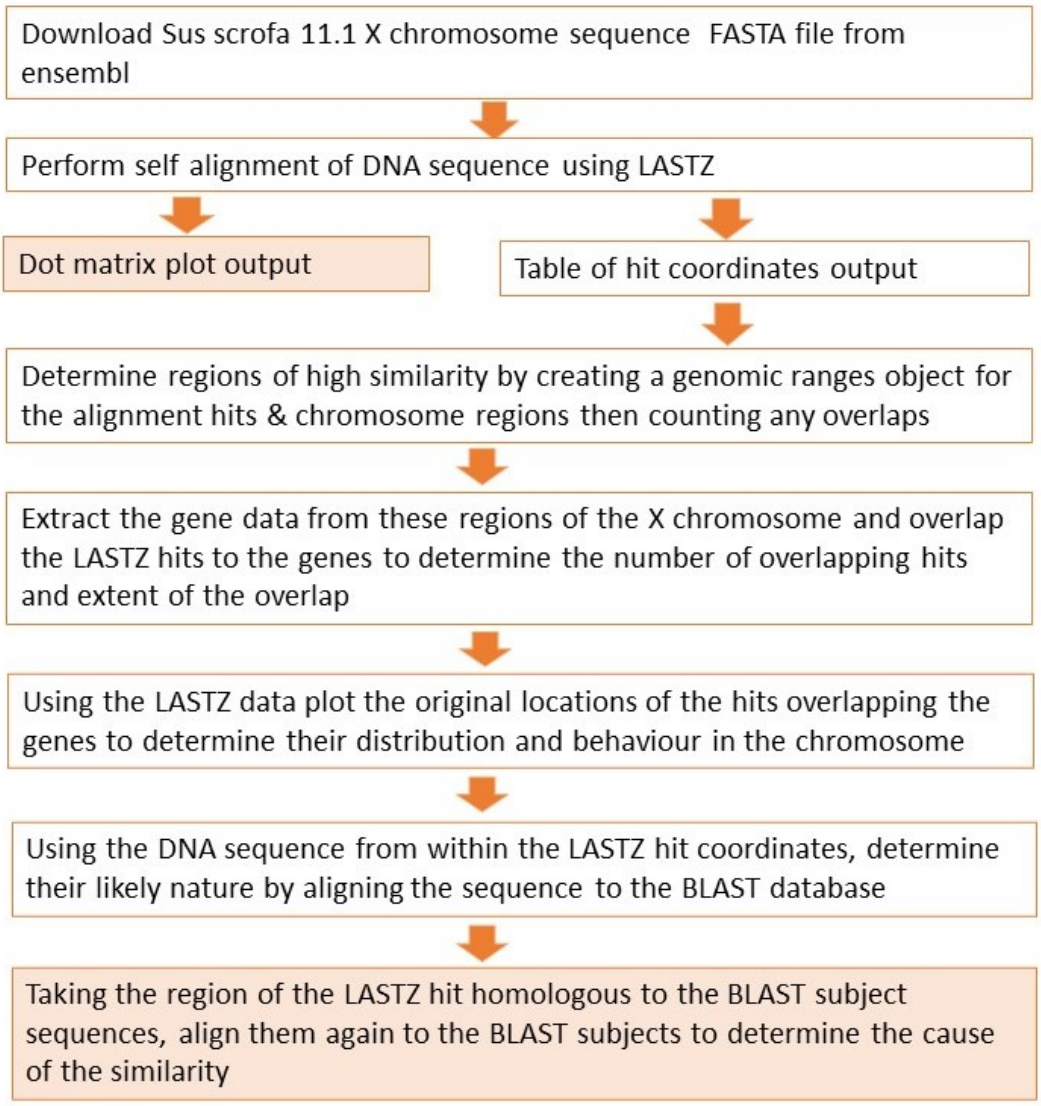

Figure 2.1 Flow chart showing the stages of the LASTZ analysis to determine highly similar regions in the pig X chromosome and identify their likely nature using BLAST. These stages were performed in the sequence shown as each stage informed which alignment hits and genes to investigate further meaning the results of each stage determined the input for the next. The shaded boxes show the final stage of the analysis, where the dot matrix plot was formed to create a general overview and was not used as further input for later analysis, and the final stage of the overall analysis involved determining the nature of the overlapping hits.

IV. Initial self-alignment of pig X chromosome using LASTZ

The FASTA DNA files of the pig X chromosome sequence were aligned to themselves using the pairwise aligner LASTZ, version 1.04.06 (Harris, 2007). The sequence acted as both the target and query input. Subsequently, the files were parsed into a seed word position table to conserve memory and runtime (see LASTZ documentation: http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html#ex_stages).

Following this, short near-matches between the target and query were identified as regions of homology known as seeds. From this stage onwards the seeds were manipulated as required through the use of different flags passed to the alignment script. In our analysis, the seeds required extension to form part of a high-scoring segment pair (HSP). This was due to the high computational demand posed by the large size of the pig X chromosome. Forming HSPs allows for a given extension rule (in our analysis the x-drop rule was employed) to determine the 'survival' of the HSPs in an alignment, for instance if a HSP does not meet the minimum HSP threshold it will be removed from the alignment (Harris, 2007).

Surviving HSPs were chained together and extended once again via gapped extension. This extension occurs in both directions from an anchor point, which is most likely to lie on the optimal path, in other words the centre of the highest scoring region of the HSP (Harris, 2010). Finally there were some extra filtering processes applied such as filtering according to the percentage identity between the query and target sequence (Harris, 2010).

The LASTZ alignment was performed on the command line and required installation as a package, directions on how to do this are found in the documentation: (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html#install). Once installed, the alignment can be performed by calling the function "lastz" and following the syntax used in the scripts found at: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/LASTZ_alignment.

Performing the self-alignment required the pig X chromosome sequence to be passed to the command using the "self" parameter flag. Self-alignments provide trivial self-alignment diagonals and mirrored alignments which were removed using the "notrivial" and "nomirror" flags. The high computational demand of the pig X chromosome led to the removal of alignments not meeting the given HSP threshold. This threshold set too high would leave no alignments

remaining, and too low would return errors due to the persisting computational demand (Harris, 2010). Through employing the preliminary low sensitivity adaptive threshold alignment; we were able to determine the lowest threshold which would return the highest number of alignments within the computational limits.

Comparing the optimal thresholds revealed a discrepancy between the unmasked and repeat-masked DNA sequences. The optimal score for the unmasked DNA sequence (31250) returned few repeat-masked alignments and lowering this score created too high a computational demand for the unmasked sequence. Considering these factors, the threshold was set at 31250 for the unmasked DNA to reduce computational limitations and the hspthresh flag was omitted for the masked sequence to allow for all alignments to be returned. For the purposes of this study the largest number of alignments are desired to increase the probability of finding ampliconic genes, therefore, this appeared to be the ideal compromise.

At this stage, there is a large quantity of 'noise' in the alignment data due to low similarity alignments. These were filtered out using the 'identity' flag to remove all hits below 95% and 99% similarity - it is known that multicopy genes often share over 95% sequence identity and ampliconic genes share over 99% sequence identity. This produced all the alignment hits of interest for our study.

Finally, two output formats were used: 1) "rdotplot" provided a naive dot matrix plot to visualise the alignments, and 2) the output table provided the start and stop positions of the alignment hits, the strand where the hits were located, and the percentage identity between the hits. This table format was required for the following analysis stages:

V. Rearrangement of the X chromosome alignment to determine regions of significantly high similarities

At this stage, the naïve dot plot showed the density of the hits on the chromosome however with very little detail. Therefore, it became necessary to rearrange the data to provide an in-depth analysis of the LASTZ alignment hits. Through the use of the GenomicRanges package in R Bioconductor version: Release (3.13), the data was rearranged to show the number of homologous alignment hits within regions of the chromosome (Lawrence et al., 2013). For ease of analysis, functions were created to perform tasks which would be repeated; these functions are in scripts seen here: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/functions.

To determine the number of overlaps between regions of the chromosome and the LASTZ hits; genomic ranges objects were created for the hits and the chromosome. A range is defined as an organised set of successive integers, in this case the integers are the start and stop positions of the target hits, where start ≤ stop (Lawrence et al., 2013). The chromosome ranges were 3000bp in length between 1bp (the chromosome start position) to 126Mbp (the chromosome end position). This window size was chosen as it encompassed the average length of the LASTZ hits. The start and stop positions taken from the LASTZ table were used to create the ranges for the hits.

The genomic ranges objects were overlapped with the function countOverlaps() which detects overlaps between GenomicRanges objects using an interval tree algorithm (Lawrence et al., 2013). Counting ‘any’ overlap in this analysis was the most suitable to visualise the extent of the similarities to find potential ampliconic genes.

The two GenomicRanges objects were overlapped to determine the number of LASTZ hits overlapping the chromosome windows. This provided a numerical matrix which was used to

visualise the number of hits overlapping the regions of the chromosome via ggplot. The plot then distinguished regions of the chromosome with significantly more LASTZ hits showing similarity.

For the following analysis the entire chromosome was investigated, with the exception of the first 7Mbp which has previously been defined as the PAR which is not of interest in this study as it is known to harbour homologous XY genes and some orthologues with other species (Skinner et al., 2013).

VI. Determining where the alignment hits show similarity in the X chromosome genes

The content of regions with high numbers of hits was investigated using the biomaRt package in R Bioconductor version: Release (3.13) (Durinck et al., 2009). All the known genes from the pig X chromosome were downloaded to aid in identifying which genes had overlapping LASTZ hits. This was done using the fetch.x.genes() function shown here: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/functions. Any genes with no overlapping LASTZ hits were discarded from the investigation, while genes with overlapping hits were passed to the next stages of the analysis.

Overlapping LASTZ hits for each gene were extracted using the findOverlaps() function and visualised using ggplot. The ggbio package Bioconductor version: Release (3.13) allowed for the gene model to be annotated with data from genome browsers such as Ensembl (Yin et al., 2012). The gene tracks created using ggbio show the introns, exons, and untranslated regions of the gene. The originating location of each hit to a gene was also visualised on an ideogram (view https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/blob/main/LASTZ_analysis/overlap.hits.entire.chromosome.R).

VII. Comparison of the alignment hit sequences to known sequences using BLAST

Any sequences not mapping to known genes were investigated using BLAST. The LASTZ hit DNA sequence (the query) was passed, in FASTA format, to NCBI BLAST (Madden, 2003) which searched for similar sequences stored within a database, this was done as a heuristic short match alignment providing statistical information about the alignment (Johnson et al., 2008). FASTA sequences were created from the hit coordinates from LASTZ using the Ensembl Perl API, via the following script: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/LASTZ_analysis. The Perl script connects to the Ensembl API and extracts the FASTA DNA sequence between the given start and stop coordinates.

To facilitate the analysis, a local copy of the nucleotide (nt) database was downloaded (26/05/21). The syntax used in the BLAST search can be found here: https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/BLAST_analysis.

The results, in tabular format, allowed commonly matching elements to be detected. Determining the most frequently occurring subject sequences in the BLAST output

The BLAST tables returned many sequences, many of which were BAC or other clone sequences which were removed as they were not of interest in the study. The accession numbers for remaining sequences were extracted and analysed in R to identify the most frequently occurring accession numbers (https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/tree/main/BLAST_analysis)

VIII. Investigating the nature of the BLAST subject sequences in the alignment hits

In many cases the BLAST results returned several different subject sequences for each alignment hit. Some of these returned subject sequences covered different regions of the alignment hit; or numerous genomic elements were homologous to one hit. The most frequently occurring

subject accessions from the frequency table (section 2.VII) were examined, and their corresponding X-chromosomal locations were visualised to determine the nature of the alignment (view https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/blob/main/BLAST_analysis/individual.blast.sh and https://github.com/Squigley18/Pig-X-chromosome-repetitive-landcape/blob/main/BLAST_analysis/blast.subject.hits.aligned.to.subject.accessions.sh).

3. Results

I. *Sus scrofa* inverted repeat content appears to be insignificant when identifying ampliconic genes

Large inverted repeats have been associated with chromosomal rearrangements, gene duplications, deletions, and amplifications (Alamro et al., 2021). Ampliconic genes have been observed within inverted repeat structures, such as in the human sex chromosomes (Lange et al., 2013; Warburton et al., 2004). With this in mind, the search for ampliconic genes within the pig X chromosome was enhanced with evaluations of the inverted repeat content of the chromosome. Using the unmasked and hard repeat-masked X sequence from Sscrofa11.1, the inverted repeat structure was investigated using the IUPACPAL software (see 2.II).

The pig X chromosome assembly is comparable to that of the human and mouse assemblies, however, it is still incomplete, shown with unassembled regions of the sequence represented using N's (Soh et al., 2014; Warr et al., 2020). This incomplete assembly and the nature of hard repeat-masking containing N's within the sequence; led to the IUPACPAL software identifying these regions as inverted repeats. Therefore, N masked sequences were removed from the data to reduce unnecessary alignments being investigated.

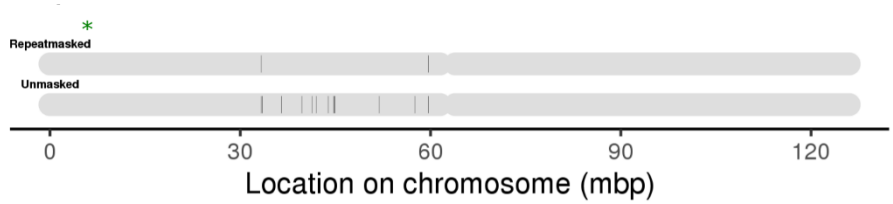

Figure 3.1 Ideogram distribution of palindromes within the pig X chromosome. The DNA sequences investigated were hard Repeat-masked (top ideogram) and unmasked (bottom ideogram). The hard-masked hits are a subset of the unmasked hits where all but 2 hits were removed due to masking. The original unmasked DNA sequence had 12 detected inverted repeats. All detected palindromes were located between 30-60Mbp on the chromosome. The PAR ranges from 0 to the PAR boundary marked with the green asterix.

IUPACPAL identified twelve inverted repeats (Figure 3.1) between the 30-60Mbp region (Table 3.1) and ranged between 50-120bp in length (Table 3.1). When hard repeat-masking was applied only 2 of these inverted repeats remained. The size and large quantity of the hits and having

been removed through repeat-masking may suggest they are of little significance where hard repeat-masking often removes low-complexity DNA. However, to determine the functional significance of these hits, the start and stop positions were compared to known gene locations to identify overlaps.

Table 3.1 Precise coordinates of palindromes detected using the inverted repeat detection tool IUPACPAL. Twelve palindromes were detected and when hard repeat-masked only two palindromes remained. The unmasked palindromes have been highlighted in blue and the hard-masked palindromes highlighted in orange.

Masking properties	Start	Stop	Arm Length (bp)	Start2	Stop2	Arm Length (bp)	Gap length (bp)
Unmasked	33455280	33455354	74	33455390	33455464	74	36
Unmasked	36499283	36499374	91	36499437	36499528	91	63
Unmasked	41941849	41941925	76	41941960	41942036	76	35
Unmasked	44840978	44841043	65	44841114	44841179	65	71
Unmasked	57487855	57487938	83	57487980	57488063	83	42
Unmasked	33302345	33302400	55	33302491	33302546	55	91
Unmasked	39684633	39684711	78	39684728	39684806	78	17
Unmasked	41337272	41337336	64	41337376	41337440	64	40
Unmasked	43834907	43835026	119	43835088	43835207	119	62
Unmasked	44733169	44733246	77	44733274	44733351	77	28
Unmasked	51905590	51905646	56	51905727	51905783	56	81
Unmasked	59619230	59619292	62	59619332	59619394	62	40
Hard Repeat-masked	33302348	33302403	55	33302494	33302549	55	91
Hard Repeat-masked	59619233	59619295	62	59619335	59619397	62	40

Four genes were identified as having overlapping palindromes (Figure 3.2), located at around 33425kbp, 41250kbp, 41950kbp and 51840kbp. Only the unmasked dataset showed to have any overlap to X chromosome genes - *SLC9A7*, *MSN*, *LANCL3*, and *ZNF41*. However, in addition to being short, all the palindromes were found within the gene introns; except for *MSN* where the overlap of the palindrome was predominantly to the intron with some overlap to an untranslated region.

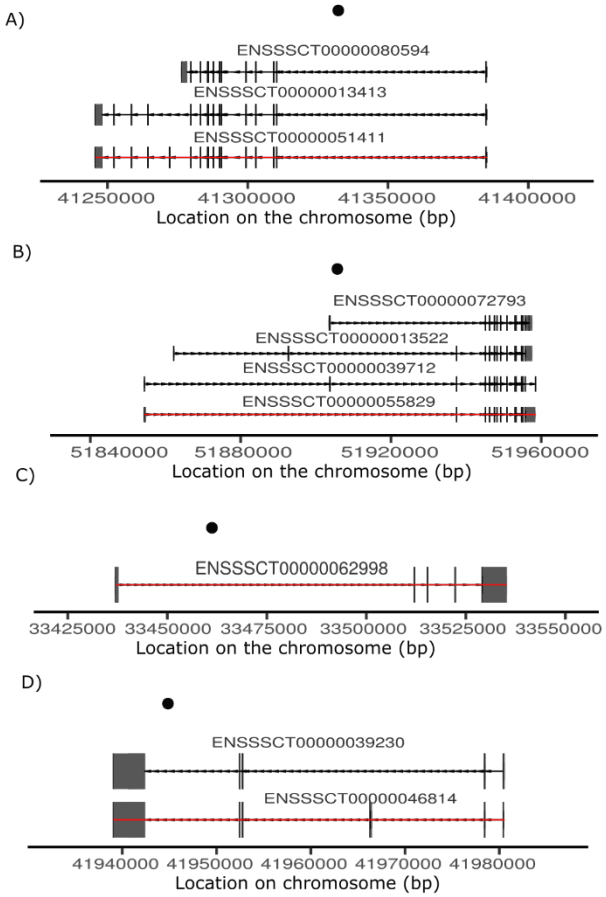

Figure 3.2 Palindromes overlapping genes of the pig X chromosome found within the unmasked DNA sequence as shown in black dots above the annotated gene (highlighted in red). The Annotations show regions where there are exons (large grey boxes), regions of introns (black arrows), and untranslated regions (small grey boxes). The detected palindromes all overlapped intronic regions of the genes. A) palindromes overlapping SLC9A7 B) palindromes overlapping MSN C) palindromes overlapping LANCL3 D) palindromes overlapping ZNF41

II. Low sensitivity LASTZ alignment suggests presence of large duplications and possible inversions

IUPACPAL detected small, inverted repeats below 120bp within introns or between genes of the X chromosome. Given no significant small palindromic repeats were detected with IUPACPAL we decided to determine if there were any significant repeats using the alignment software LASTZ. To determine if any regions were outstanding with large repeats, we ran a low sensitivity alignment.

The alignments showed there to be some regions of interest in the 99% identity repeat-masked dataset; therefore, we ‘zoomed in’ on these regions as shown in Figure 3.3. We determined there

were some inversions and duplications of interest in these regions, and they have been investigated further in our LASTZ analysis (section 3.V onwards).

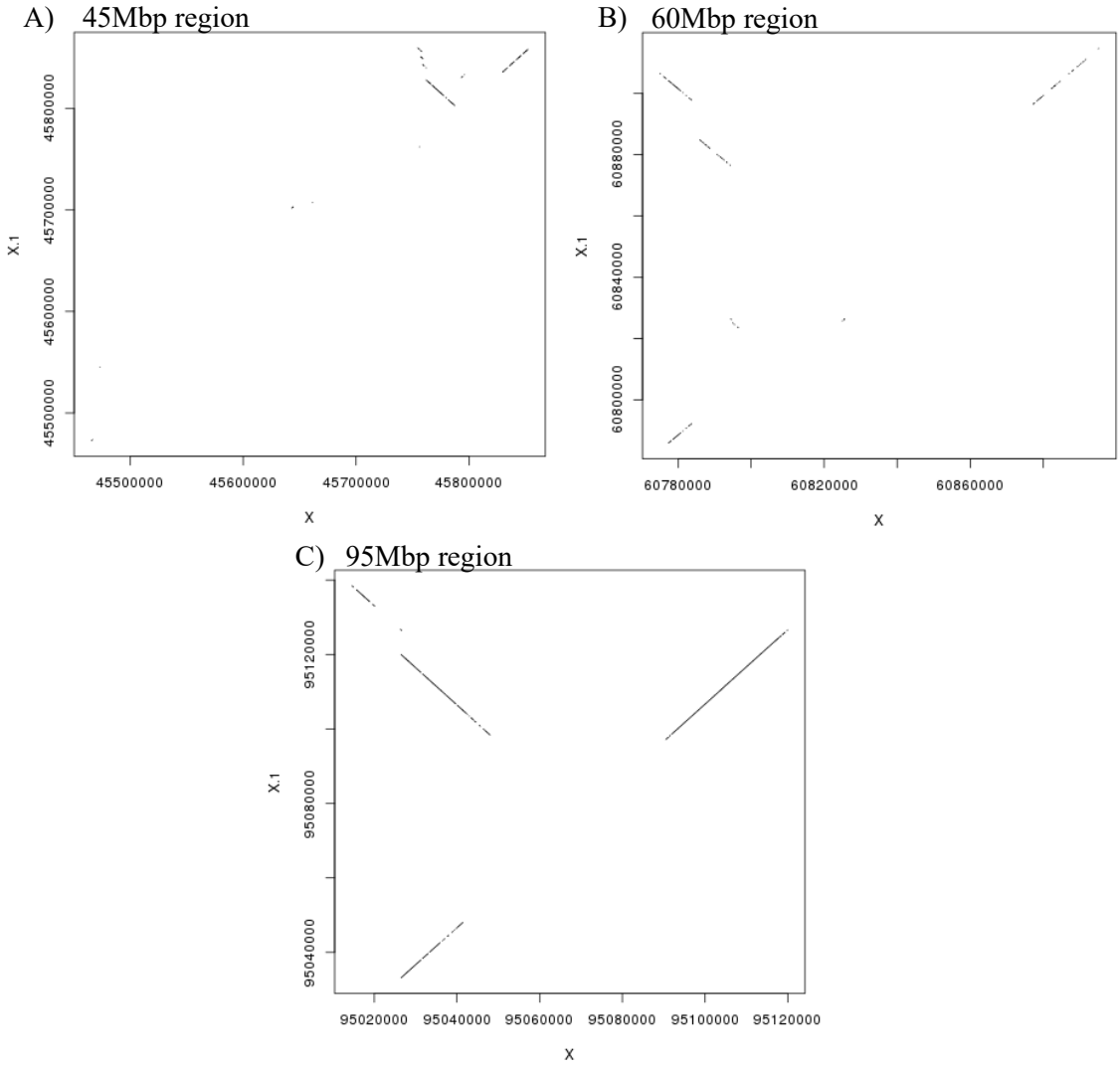

Figure 3.3 Low sensitivity self-alignments of the pig X chromosome to itself filtered at 99% with hard Repeat-masking applied to the DNA sequence. These graphs show 'zoomed in' regions of the chromosome which suggested there to be content of interest in the large-scale alignment graphs. A) around the 45Mbp section of the chromosome. B) around the 60Mbp region of the chromosome. C) around the 95Mbp region of the chromosome. These graphs show alignment hits suggesting likely duplications and inversions have occurred.

III. High density of self-homology along the entire pig X chromosome apparent in naïve self-alignment

IUPACPAL showed no significant small palindromic repeats to be present within the Sscrofa11.1 X assembly, however, the low sensitivity LASTZ alignment suggested some regions

may contain large duplications and inversions. To investigate these further and to determine if there are other duplicated or ampliconic sequences present in the pig X chromosome; we used the alignment software LASTZ to fully align the X chromosome sequence against itself. We performed the following alignments and filtering of the outputs: 1) the complete X sequence; 2) the hard repeat-masked X sequence. This allowed us to assess the degree of repetitive content and identify previously unannotated elements.

Multicopy and ampliconic genes share significant homology between copies, therefore, we filtered the alignment hits according to their percentage similarity 1) above 95% and 2) above 99%. These values were chosen due to the known nature of ampliconic genes to be 99% identical (Bellott & Page, 2009; Skaletsky et al., 2003); and multicopy genes to be around 95-97% identical (Sudmant et al., 2010).

The pig X chromosome is dense with self-homology suggesting the presence of repetitive elements (and potentially ampliconic genes). These regions of homology are significantly similar where the alignments have been filtered to remain above 95% identity (Figure 3.4A&C) and above 99% identity (Figure 3.4B&D). The large number of alignment hits remaining after this filtering suggested there may be ampliconic genes present in the pig X chromosome. However there are also a large number of known repetitive elements within genomes such as retrotransposons and long terminal repeats which also have high percentage identity and are often found in high density in mammalian X chromosomes (Groenen et al., 2012; Le Rouzic & Capy, 2005). These known repetitive elements were removed from the pig X chromosome sequence using RepeatMasker (Smit et al., 1999) (Figure 3.4C&D). As a result of the filtering and hard repeat-masking there was a significant reduction in noise from low identity and repetitive elements, yet still there remained an abundance of alignment hits across the chromosome. This further suggested the presence of either amplicons or unannotated repetitive elements.

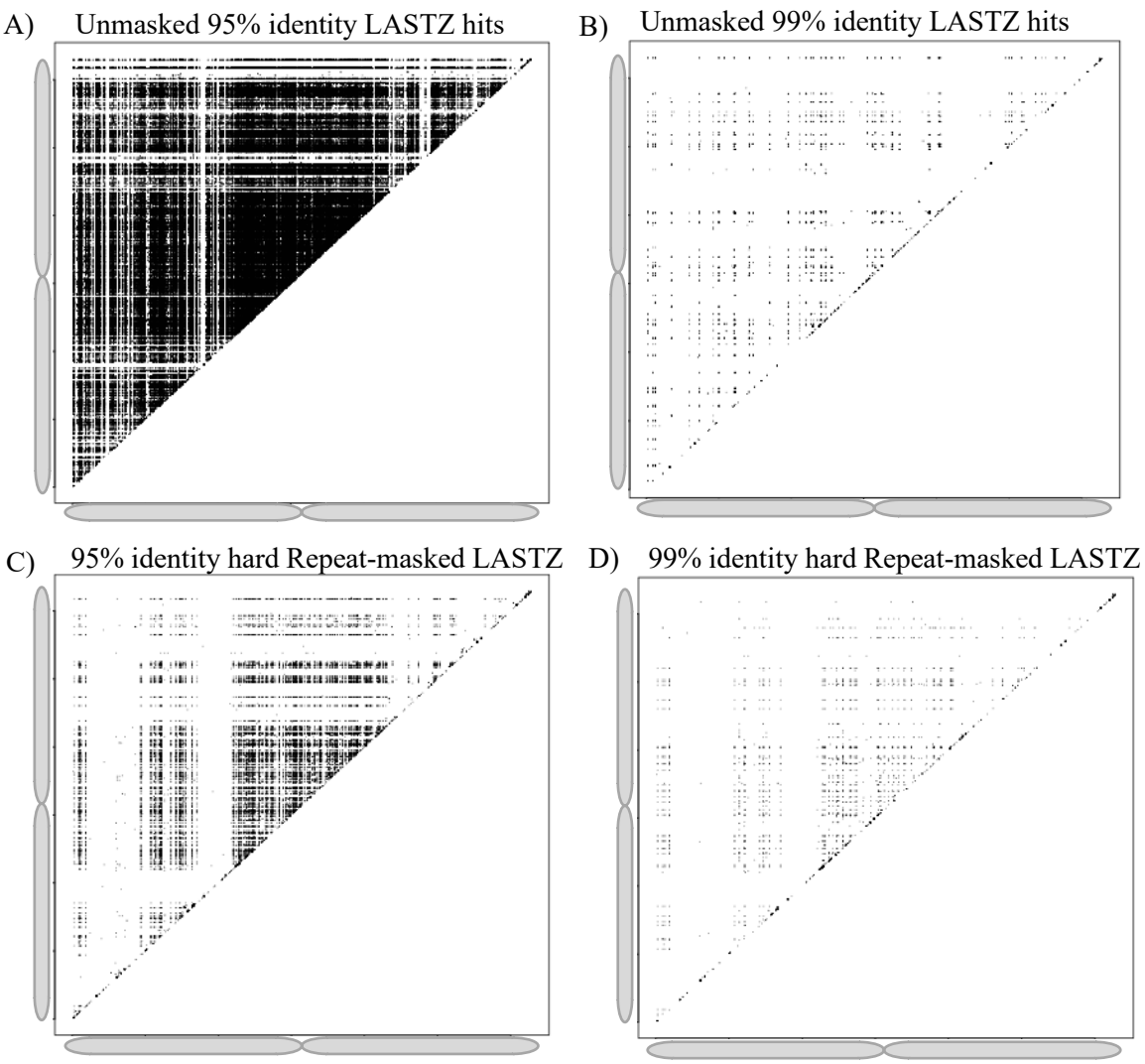

Figure 3.4 Dot matrix plot created using the raw LASTZ alignment data. The ideogram outlines as shown in grey on the X and Y axis represent the Sus scrofa X chromosome DNA sequence aligned to both axes. The LASTZ hits have been filtered to remove hits below 95% identity A&C and 99% identity B&D. There has also been hard repeat-masking applied to the DNA sequence in C&D.

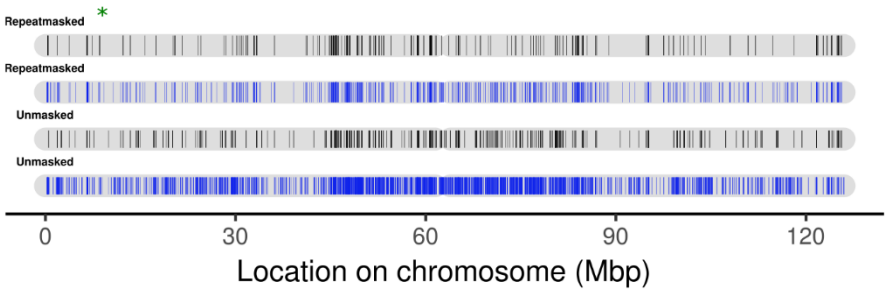

Figure 3.5 LASTZ alignment hits distribution along the chromosome ideogram. The top two ideograms show the DNA sequence which had been hard repeat-masked the bottom two ideograms shows the unmasked DNA sequence. The hits in blue represent the alignment hits having been filtered to remove alignments under 95% identity and the hits in black represent the alignments filtered to remove any hits with lower than 99% identity. The PAR ranges from 0 to the PAR boundary marked with the green asterix.

The naïve dot matrix plots (Figure 3.4) and the ideogram showing the hit distribution (Figure 3.5) showed the centromere to have the greatest density of alignment hits correlating with

known patterns of transposable elements congregating within non-recombining regions (Wright et al., 2017). Next, the pericentromere and chromosome arms annotated with numerous hits may be the result of ampliconic genes or constitutive heterochromatin which is also often 'riddled' with transposable elements (Janssen et al., 2018). The PAR (~1-7Mbp region) has a high density of hits, however, is not of interest in this study due to it being a region of sex chromosomes where XY homology has remained; ancestral DNA has yet to be degraded, and recombination still occurs. These properties of the PAR leave it unlikely that significant multicopy genes are present. At this stage the alignment hits are interpreted with little detail and therefore to further understand where these hits are found and what they contain, an in-depth analysis was performed.

IV. Regions of the Sus scrofa X chromosome harbour higher densities of self-alignment hits than others

Thus far, the Sscrofa11.1 X assembly has been shown to be dense with self-homology (Figure 3.4 & Figure 3.5). Determining the extent of this required annotating the number of hits homologous to regions (windows of 3000bp, 2.V) of the chromosome.

Some regions 'stand-out' as containing a high density of hits; these regions are consistent between the different identity and masking filtered datasets. Regions of interest have been highlighted in Figure 3.6 in grey. Interestingly, the 35-50Mbp region, 60-70Mbp region, and 95-110Mbp region correlate with the regions of interest in Figure 3.3.

In order to prioritise regions of self-homology for further investigation, we identified the types of sequence they contained, and where in the chromosome these homologous hits originated.

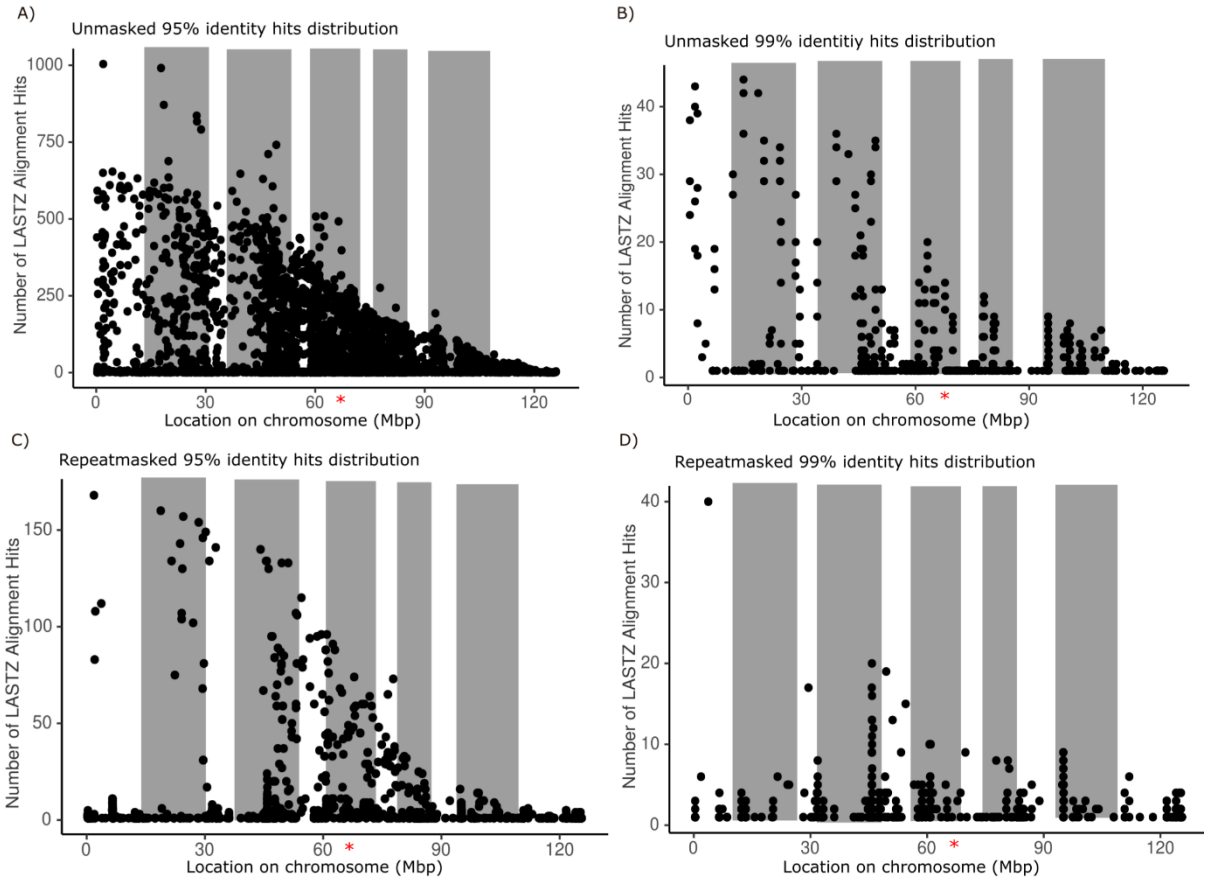

Figure 3.6 Rearrangement of LASTZ hit distribution along the pig X chromosome showing the number of hits at each region. The pig X chromosome was split into windows of 3000bp with the number of LASTZ hits overlapping each window calculated. The regions covered in grey boxes are highlighting chromosomal locations with consistently high numbers of self-alignment hits across the filtering processes. The red asterix highlights the location of the centromere. A) Unmasked 95% identity hits B) Unmasked 99% identity hits C) repeat-masked 95% identity hits D) repeat-masked 99% identity hits

V. LASTZ hits showing homology to the genes in the Sus scrofa X chromosome fall into several patterns

The extent of self-homology in the Sscrofa11.1 X assembly is apparent (Figure 3.5 & Figure 3.6); the next stage was to determine the cause of this homology. There are around 1311 known genes in the pig X chromosome to date (see Table 1.1). These genes comprise of both coding and non-coding genes as well as pseudogenes. The LASTZ hits were investigated to determine if they overlapped any known genes in the X chromosome. In total 150 genes were found with overlapping LASTZ hits from the 99% identity datasets (repeat-masked and unmasked).

Many of these genes have been included broadly within the investigation, however, not all 150 genes were investigated in depth. Segment plots (e.g Figure 3.6), a visual representation of the

homology shared between the genes and LASTZ hits, were produced for the repeat-masked and unmasked datasets at both 95% and 99% identity. Observations of all the segment plots revealed patterns which could be used to group the genes and allow for individual representatives to be selected. These patterns were relatively similar between the 95% and 99% identities; the major differences being the quantity of hits, as to be expected, as many hits were removed when the percentage identity threshold was increased. Due to the similarities the 99% identity hits were focused on for this stage.

The genes with overlapping LASTZ hits fell into 7 major groups showing the overlap patterns of the hits to the gene. The groups included: 1) genes entirely covered by the LASTZ hits (15 genes), 2) small hits stacked within regions of the gene (25 genes), 3) small hits staggered throughout the genes (3 genes), 4) LASTZ hits partially covering the genes (8 genes), 5) very small LASTZ hits overlapping genes which are barely visible, if not entirely undetectable in the graphs (32 genes), 6) genes with overlapping LASTZ hits from only the unmasked sequence (45 genes), and 7) genes with overlapping LASTZ hits from only the hard repeat-masked sequence (22 genes). The major patterns were investigated further.

LASTZ hits overlapping the entire gene from both the unmasked and repeat-masked DNA sequences show replication of genes

There were 15 genes which were entirely covered by the LASTZ hits, however, with a varied number of hits e.g.1-2 hits (Figure 3.7 Figure 3.8), 4-5 hits (Figure 3.9), or 10+ hits (Figure 3.10). Not all of the hits cover the entire gene, although, the majority do as in Figure 3.9 & Figure 3.10. Many of the hits exceed the length of the annotated genes suggesting this may be the result of a segmental duplication carrying the gene within the alignment. Another interesting feature of this pattern appears to be the location of the gene/hits in the chromosome correlating with the region of interest in Figure 3.3C around the 95-125Mbp region.

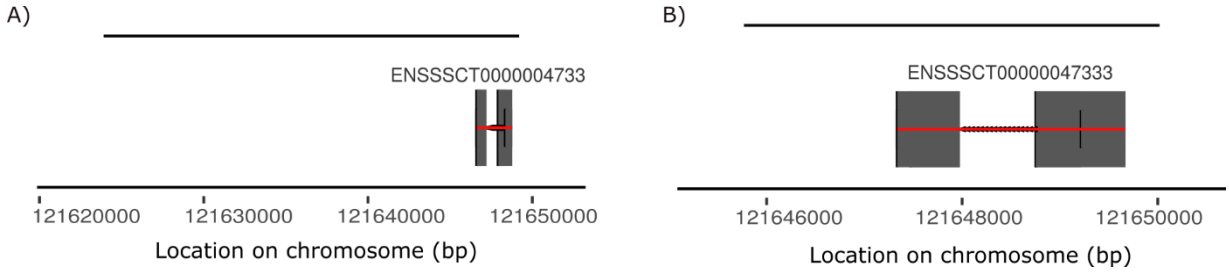

Figure 3.7 LASTZ hits showing homology to the novel gene with the Ensembl ID ENSSSCG00000040153. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) One large unmasked 99% identity LASTZ hit overlapping the entire gene B) One large repeat-masked 99% identity LASTZ hits overlapping the entire gene

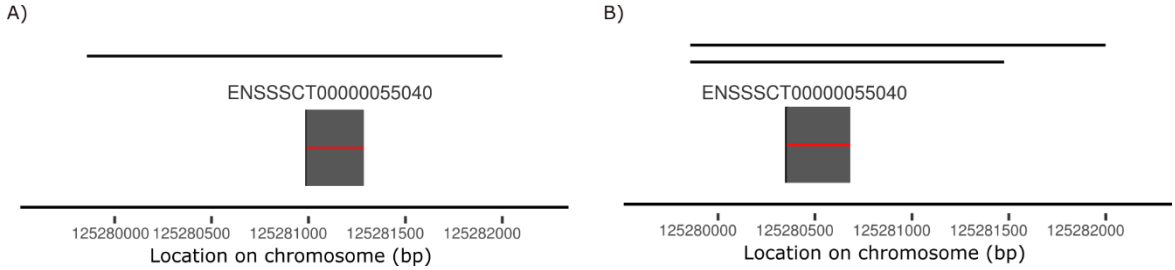

Figure 3.8 LASTZ hits showing homology to the novel gene with the Ensembl ID ENSSSCG00000034475. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) One large unmasked 99% identity LASTZ hit overlapping the entire gene B) Two large repeat-masked 99% identity LASTZ hits overlapping the entire gene

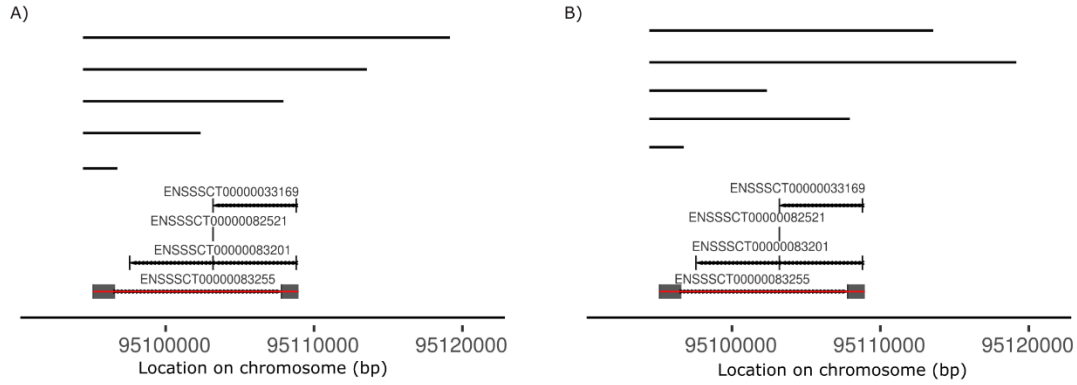

Figure 3.9 LASTZ hits showing homology to the novel gene with the Ensembl ID ENSSSCG00000048704. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) Five large unmasked 99% identity LASTZ hits with the majority overlapping the entire gene B) Five large repeat-masked 99% identity LASTZ hits with majority overlapping the entire gene

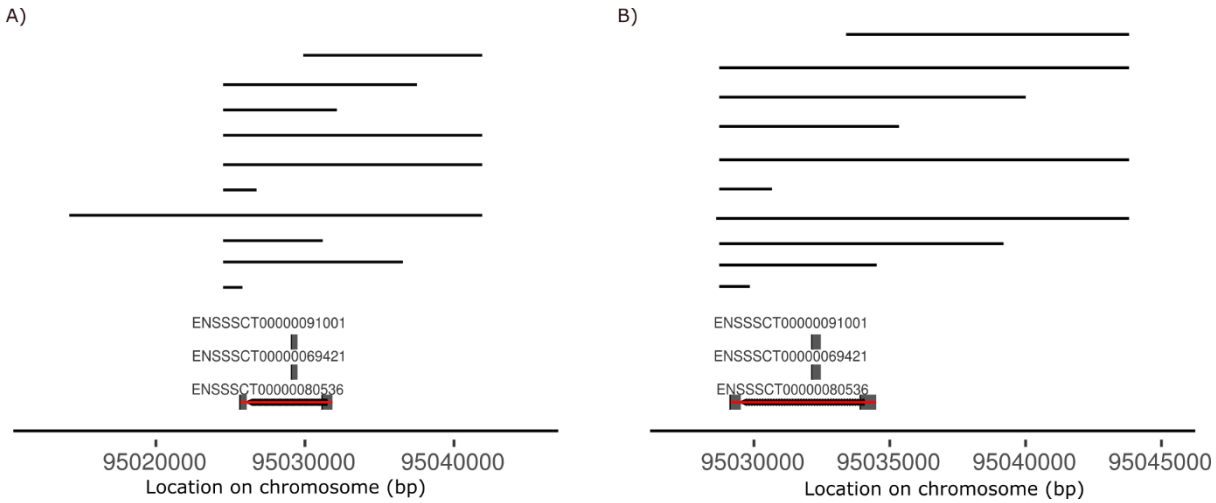

Figure 3.10 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000051484. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) Ten large unmasked 99% identity LASTZ hits with the majority overlapping the entire gene B) Ten large repeat-masked 99% identity LASTZ hits with majority overlapping the entire gene

Numerous LASTZ hits stacked within regions of the genes may be due to highly repetitive elements embedded in the gene

Instead of the entire gene being covered by the LASTZ hits, some genes have multiple smaller hits ‘stacked’ within specific regions showing a highly repetitive nature. Similarly, to the previous pattern, this has been observed in both the unmasked and masked data with various hit frequency and sizes. This pattern however has significantly more hits found within each graph compared to the other patterns observed. Many of these stacking segments such as in Figure 3.11, Figure 3.12, and Figure 3.13 are likely a result of a significant repetitive element.

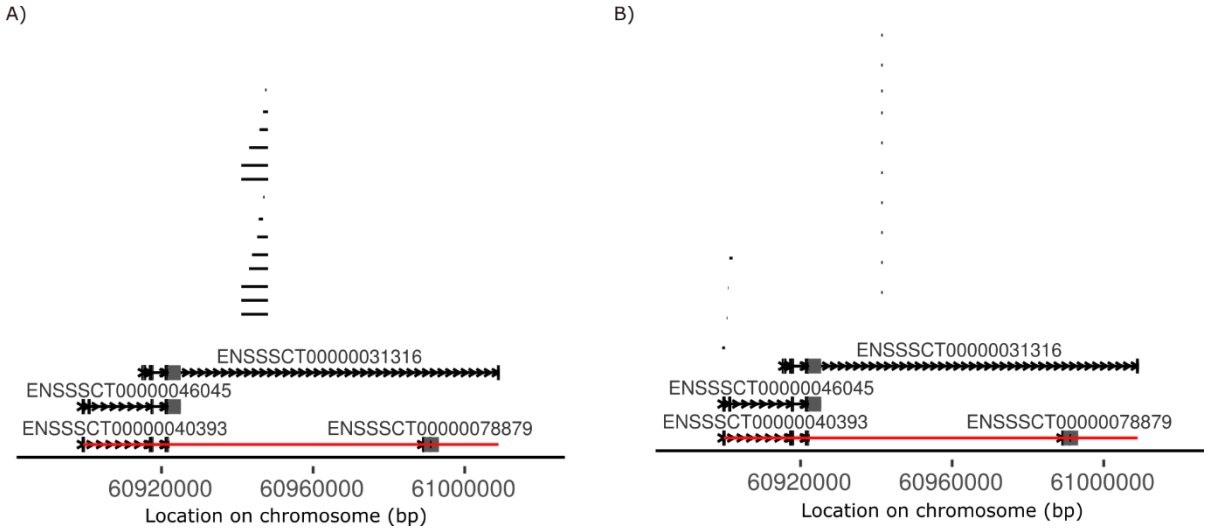

Figure 3.11 LASTZ hits showing homology to the gene PBDC1. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) 99% identity unmasked LASTZ hits stacked in one region of the gene appearing to be within an intron B) Repeat-masked 99% identity LASTZ hits stacked in one region of the gene appearing to be within an intron

Unlike the genes covered entirely by the LASTZ hits in Figure 3.7, Figure 3.8, Figure 3.9, and Figure 3.10, the genes with 'stacked' LASTZ hits are not found at discrete locations. Figure 3.14 shows the localisation of all the genes with LASTZ hits overlapping their entirety (A) and 'stacked' within them (B). (A) shows discrete clustering of the genes compared to the interspersed genes in (B). However, there are some regions showing a higher density of genes with 'stacked' LASTZ hits, particularly around the centromere (~60Mbp) which correlates with the region of interest in Figure 3.3B. Centromeres are known for their highly repetitive content including retrotransposons and long terminal repeats. Therefore, the nature of these hits appearing to be highly repetitive and many of them being found near the centromere; strongly suggests these hits to be the result of unannotated repetitive elements.

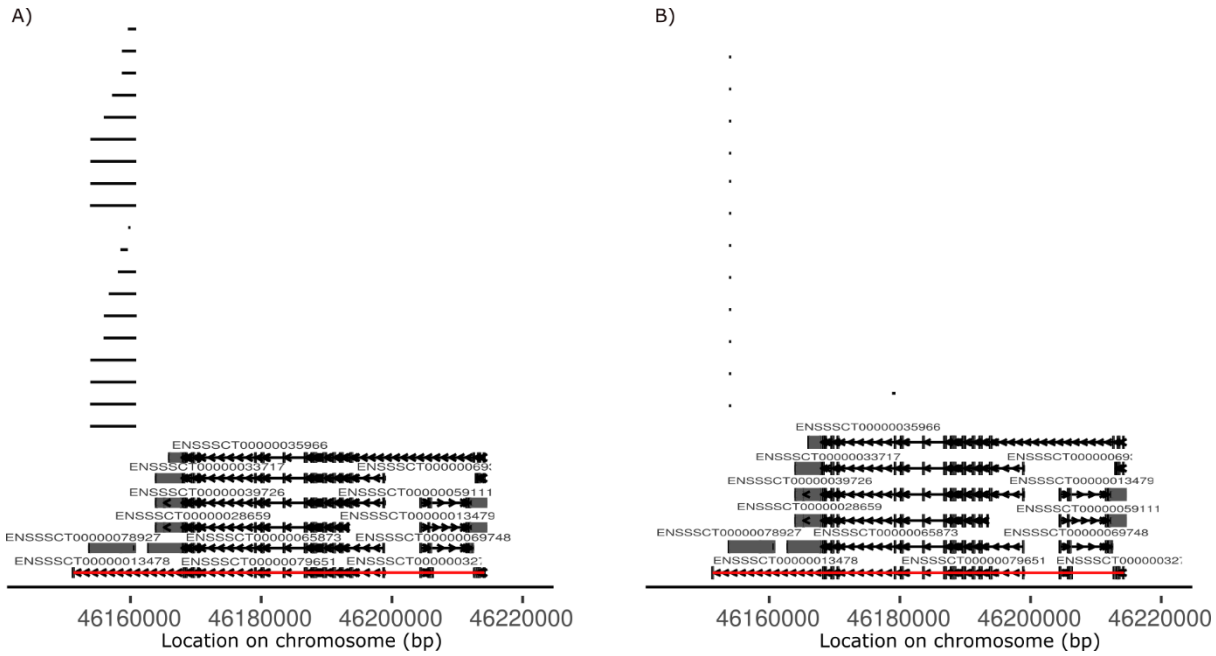

Figure 3.12 LASTZ hits showing homology to the gene SPIN3. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) 99% identity unmasked LASTZ hits stacked in one region of the gene appearing to be within an intron B) Repeat-masked 99% identity LASTZ hits stacked in regions of the gene appearing to be within an intron

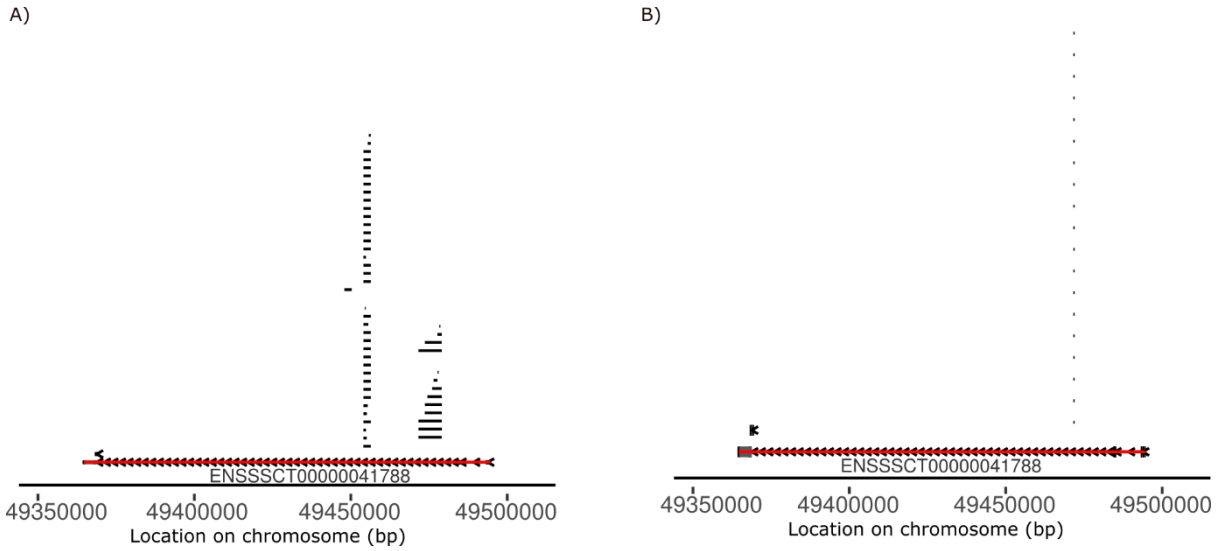

Figure 3.13 LASTZ hits showing homology to the gene SMC1A. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) 99% identity unmasked LASTZ hits stacked in one region of the gene appearing to be within an intron and partial exon coverage B) Repeat-masked 99% identity LASTZ hits stacked in regions of the gene appearing to be within an intron partial exon coverage

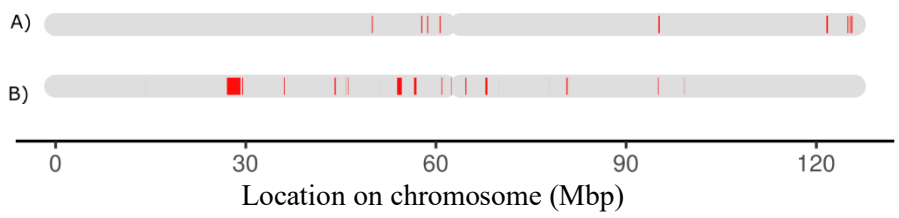

Figure 3.14 Distribution of genes (in red) with overlapping LASTZ hits. A) the genes with LASTZ hits overlapping their entirety B) the genes with smaller LASTZ hits stacked within regions of the gene

Genes with the majority of LASTZ hits partially covering them appear to show similarities with the previous patterns of hits

The previous patterns of hits have shown to either have few hits covering the entire gene or multiple small hits stacked upon one another. This next group of genes show a combination of the two. In some cases, there are one or two hits covering most of the gene but not its entirety (Figure 3.15). There are also instances of a hit covering the entire gene and many others within the same gene stacked upon each other showing only partial coverage (Figure 3.16). The large hit covering the entire gene from Figure 3.16A reduces in size with repeat-masking applied (Figure 3.16B). This may be the result of repetitive elements or low complexity DNA being removed from the alignment hits. There is no defined localisation of these genes where 6 of 8 are found in different locations (Figure 3.17).

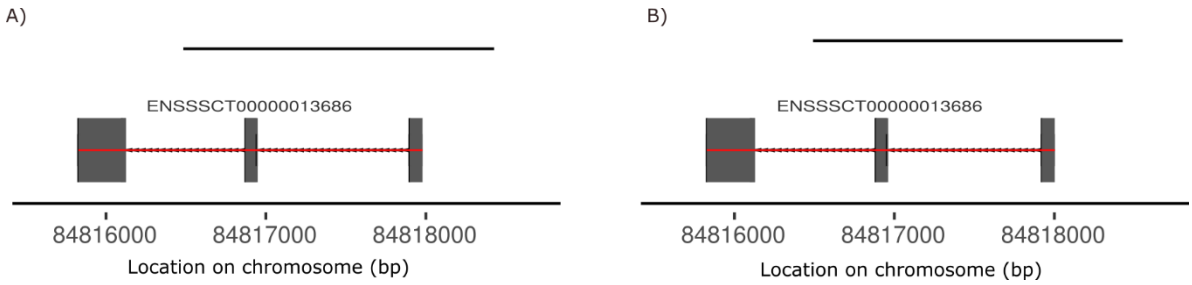

Figure 3.15 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000012517 (TMSB15B). The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) 99% identity unmasked LASTZ hits partially covering the gene with some intron and exon coverage B) 99% identity repeat-masked LASTZ hits partially covering the gene with some intron and exon coverage

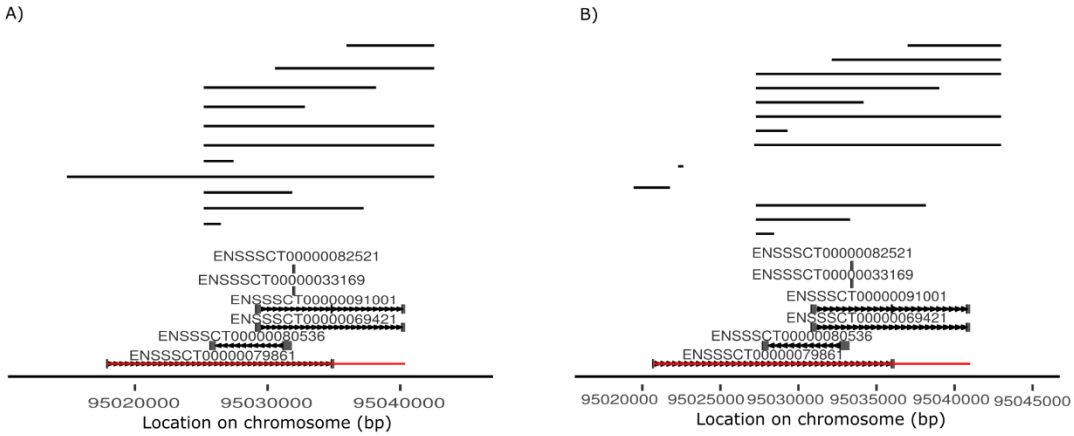

Figure 3.16 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000044950. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) 99% identity unmasked LASTZ hits partially covering the gene with some intron and exon coverage B) 99% identity repeat-masked LASTZ hits partially covering the gene with some intron and exon coverage

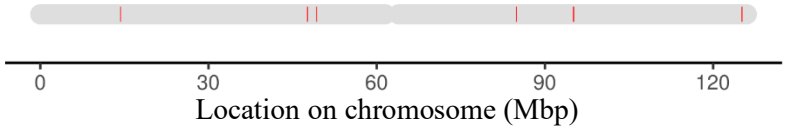

Figure 3.17 Distribution of genes (in red) with the majority of overlapping LASTZ hits partially covering the gene

Genes with unmasked hits overlapping them have no repeat-masked hits - likely due to removal of repetitive elements

The previous pattern of alignment hits has shown examples of large alignments reducing in size due to the application of repeat-masking (Figure 3.16B); this next pattern showed in some cases entire hits can be removed; likely due their being annotated as repetitive elements. The unmasked alignments have shown patterns associated with likely repetitive elements (as in Figure 3.11,Figure 3.12,Figure 3.13) however; they have also shown similarity to some of the patterns associated with likely gene duplications (as in Figure 3.8, Figure 3.9, Figure 3.10).

The removal of the stacked LASTZ hits (Figure 3.19) and the small hit (Figure 3.21) through repeat-masking would be expected as they are likely repetitive elements or low complexity DNA. Particularly the stacked alignments which have a high number of hits suggesting an embedded repeat such as a retrotransposon has been detected. The larger hits may also be repetitive

elements having been detected (Figure 3.18 Figure 3.20) however the singular hit doesn't necessarily show properties associated with highly repetitive elements. Further investigation of these hits is required to determine their nature.

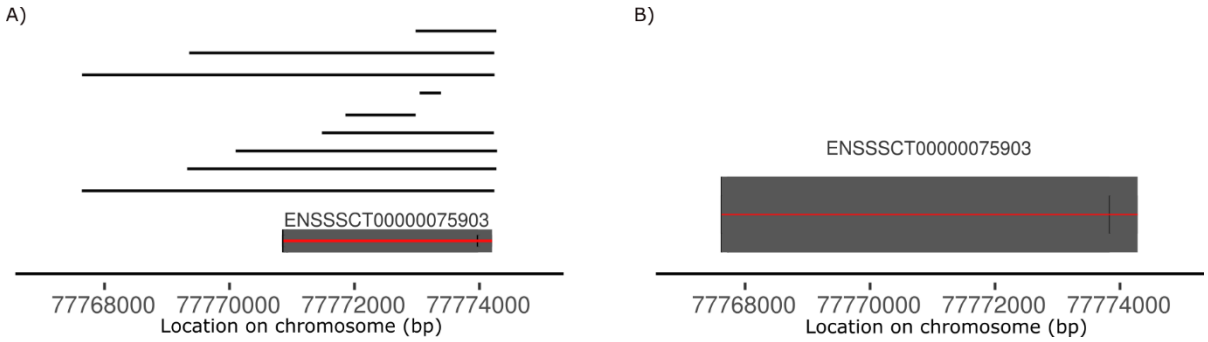

Figure 3.18 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000042737. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) 99% identity unmasked LASTZ hits stacked covering the entire gene B) Repeat-masked 99% identity LASTZ hits showing no overlapping hits

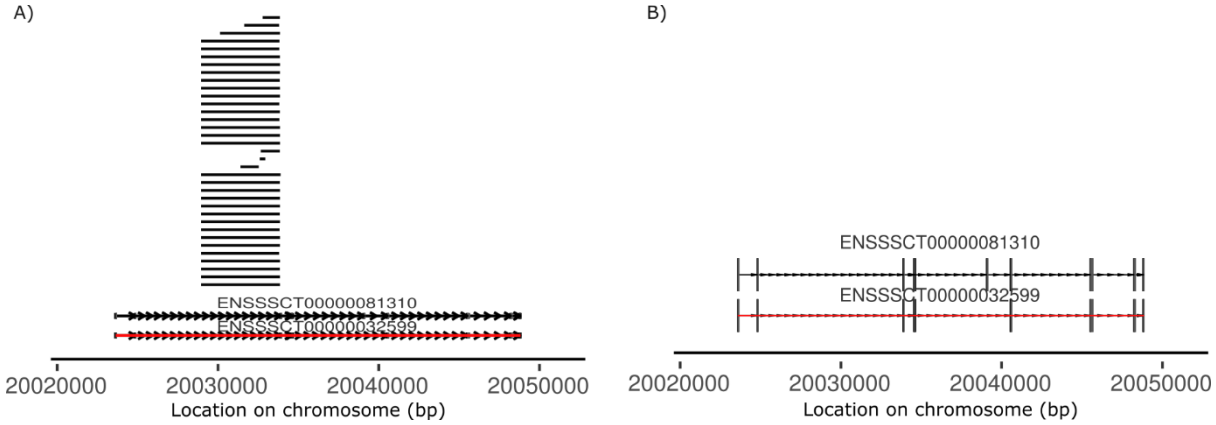

Figure 3.19 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000012175. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) 99% identity unmasked LASTZ hits stacked covering a region of the gene within an intron B) Repeat-masked 99% identity LASTZ hits showing no overlapping hits

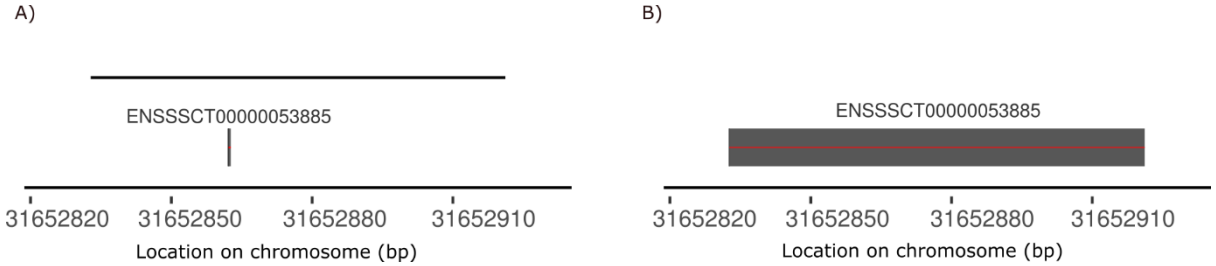

Figure 3.20 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000036038. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) One large 99% identity unmasked LASTZ hit overlapping the entire gene B) no repeat-masked 99% identity LASTZ hits

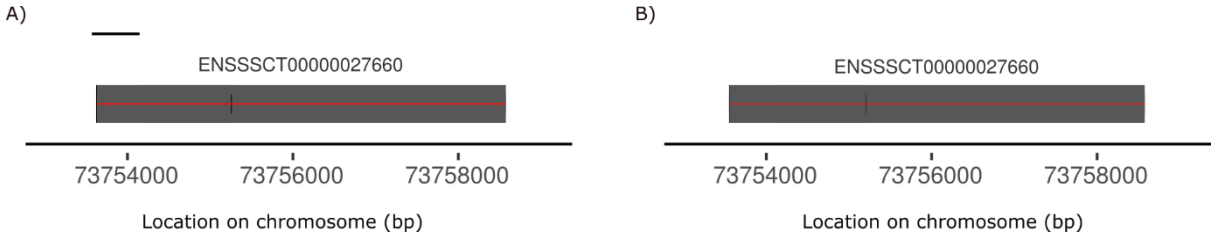

Figure 3.21 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000020695. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) One small unmasked 99% identity LASTZ hit overlapping a region of the gene exon B) no repeat-masked 99% identity LASTZ hits

Staggered hits may be the result of 'broken up' alignments covering the entire gene due to the incomplete assembly of the X chromosome

In contrast to hits covering the entire gene; some hits appear to cover the majority of the gene length however this is through numerous small hits 'staggered' through the gene. The unmasked LASTZ hits covering the genes in Figure 3.22 Figure 3.23 are larger than the repeat-masked hits suggesting there were repetitive elements which have been removed from the alignments. The genes with these overlapping hits are found within the 45Mbp region (Table 3.2). These genes are found in the regions of interest identified in Figure 3.3A. The alignments shown in Figure 3.3A confirm the suggestion that these may be larger alignment regions broken into smaller chunks.

Table 3.2 Precise locations of genes with small overlapping LASTZ hits staggered throughout the length of the gene.

Gene	Start position	Stop position
ENSSSCG00000048218	45,826,558	45,847,609
ENSSSCG00000042077	45,839,956	45,856,710

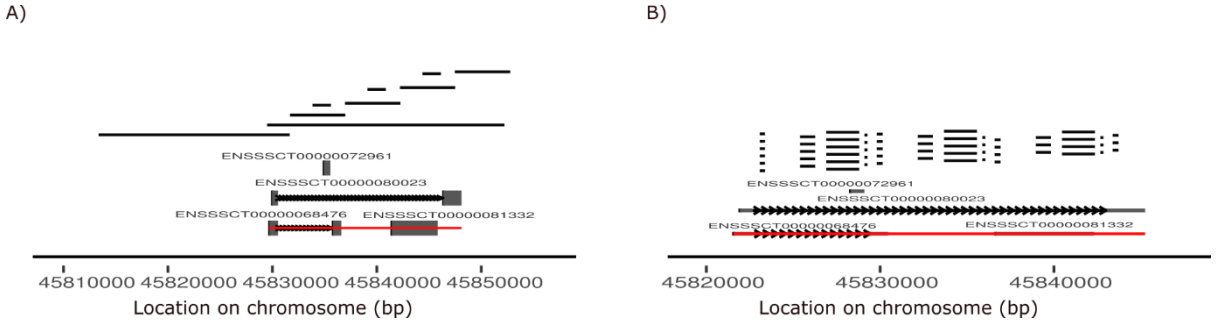

Figure 3.22 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000048218. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) 99% identity unmasked LASTZ hits staggered throughout the gene with some intron and exon coverage B) Repeat-masked 99% identity LASTZ hits staggered throughout the gene with some intron and exon coverage

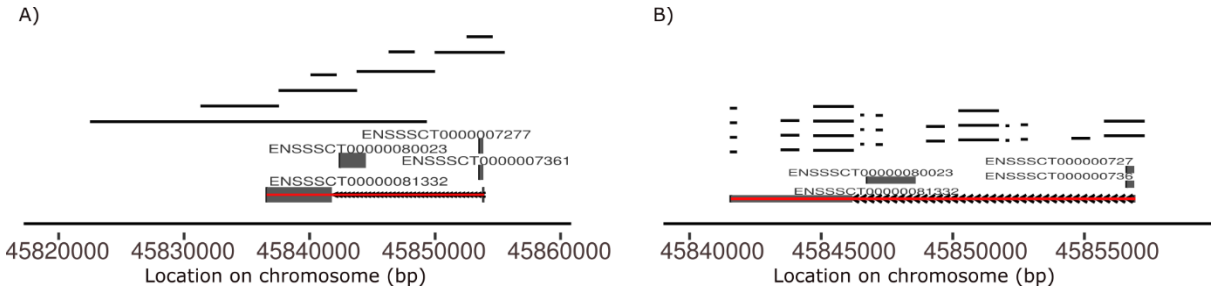

Figure 3.23 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000042077. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) 99% identity LASTZ unmasked hits staggered throughout the gene with some intron and exon coverage B) Repeat-masked 99% identity LASTZ hits staggered throughout the gene with some intron and exon coverage

Genes with overlapping repeat-masked hits have no unmasked hits - likely due to the filtering parameters required during the LASTZ alignment

It has been previously observed that some genes have overlapping hits from the unmasked data and not from the repeat-masked data (e.g., Figure 3.19). This appears to be due to the removal of repetitive elements, however, the reverse is also true (Figure 3.24). Where there have been 99% identity repeat-masked LASTZ hits overlapping genes and no 99% identity unmasked hits; is a likely result of the filtering requirements in the LASTZ programme. As explained in section 2.IV the hspthresh flag was required to lower the computational demand when aligning the unmasked DNA sequence to itself. This led to alignments not meeting the score threshold being dropped.

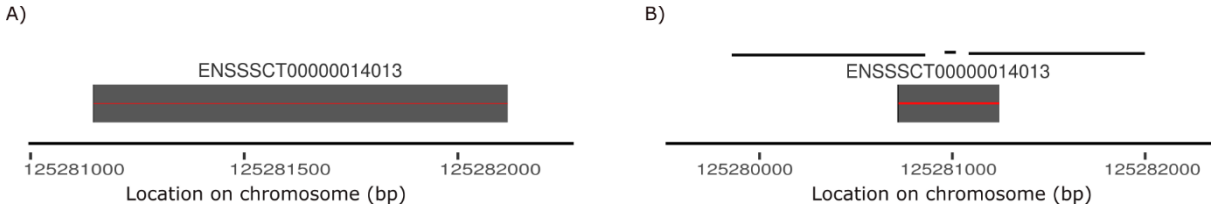

Figure 3.24 LASTZ hits showing homology to the gene with the Ensembl ID ENSSSCG00000031635. The LASTZ hits overlapping the genes were plot as segments (in black) overlapping the gene (in red). These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene A) no 99% identity unmasked LASTZ hits B) three repeat-masked 99% identity LASTZ partially covering the genes exon

Many of the LASTZ hits were too small to be detected in the segment graphs and are likely a result of low complexity DNA

The previous patterns had clear overlaps, however in this next group many of the hits were undetectable in the graphs due to their small size. To combat this the hits were plotted as dots rather than segments to identify where they were found within the gene. Many of these graphs showed to have one to two hits aligned to the introns or untranslated regions of the gene (Figure 3.25 and Figure 3.26). The small size of the hits and the regions of overlap suggest these hits are the result of low complexity DNA. This is also likely where some alignments were dropped in the score threshold filtering where there are no hits in the unmasked data but found in the repeat-masked data (Figure 3.26). This led to the decision to not focus further on these LASTZ alignments.

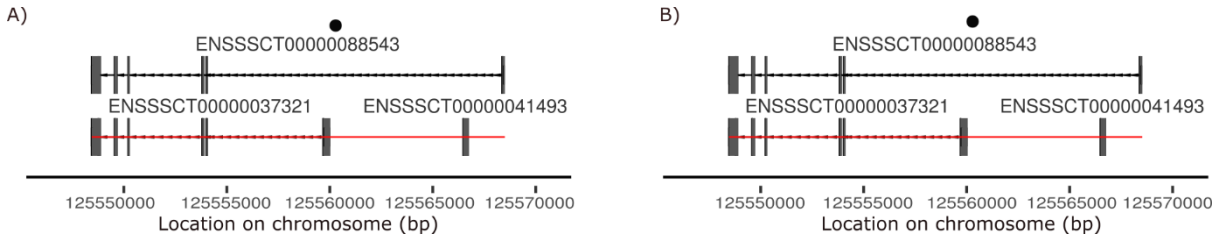

Figure 3.25 Alignments of LASTZ hits to the gene with the Ensembl ID ENSSSCG00000039998 plot as dots as the segments were not visible in the plot. These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) 99% identity unmasked LASTZ hits found within the intron B) Repeat-masked 99% identity LASTZ hits found within the intron

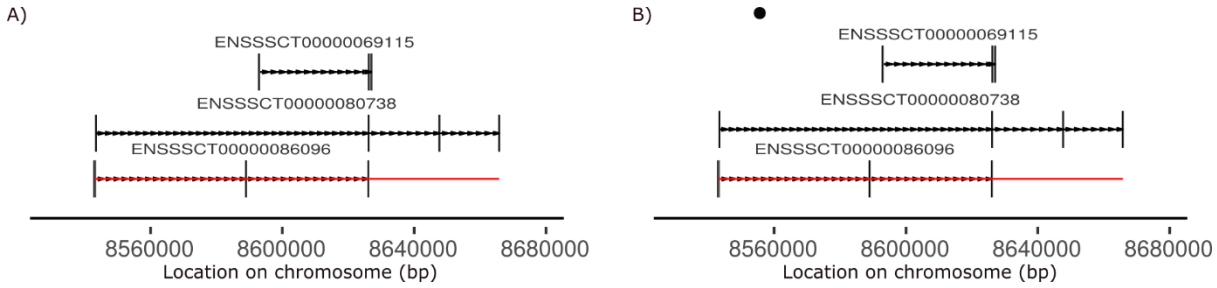

Figure 3.26 Alignments of LASTZ hits to the gene with the Ensembl ID ENSSSCG00000048531 plot as dots as the segments were not visible in the plot. These plots were also annotated using ggbio, where the large grey boxes represent gene exons, the black arrows represent the introns, and the smaller grey boxes are untranslated regions of the gene. A) No 99% identity unmasked LASTZ hits B) Repeat-masked 99% identity LASTZ hits found within the intron

VI. Distribution patterns of LASTZ alignment hits throughout chromosome suggest many of the hits to be repetitive elements or duplicated genes

Segment graphs are useful to show the extent of homologous hits overlapping genes in the chromosome; however, there is little context such as, their original location of the hits and their distribution patterns. Understanding these distribution pattens provides an insight to the likely nature of the hits due to their behaviours. Representing this distribution in an understandable and informative way is important; and therefore, the hits were plot as vertical segments onto an ideogram of the pig X chromosome.

The hits were divided into the alignment patterns detailed above (section 3.V); in short, the coverage of the gene by the LASTZ hits whether entirely, partially, only unmasked hits, only repeat-masked hits, very small hits, staggered hits, or stacked hits. When assessing the distribution of these hits throughout the chromosome; there were also some notable patterns corresponding with these segment groups. The following ideograms show the original distribution of the LASTZ hits which showed homology to the genes in the segment graphs.

The LASTZ hits which had overlapped the entire gene are found in discrete locations along the X chromosome suggesting potential gene duplications occurred

The first pattern observed in the segment plots were the LASTZ hits covering the entirety of the gene, for both the unmasked and repeat-masked data. The distribution of these hits throughout the chromosome showed the hits to originate from discrete locations. The genes homologous to these hits were located at 95-125Mbp (Figure 3.14A); with the hits originating from the same region as the gene and another location.

For example, the gene located at around 125Mbp (Figure 3.27A, red) has hits originating both at 125Mbp and within the PAR (Figure 3.27A, black & blue). Origin of the hit in the PAR may show a duplicated ancestral gene where PAR genes are regions of XY homology and have

orthologues with other species (Gil-Fernández et al., 2020). The 95Mbp gene also appears to have duplicates at 30Mbp (Figure 3.27C) which is or interest as the 95Mbp region was of interest showing a likely duplication/inversion (Figure 3.3 C). The gene in Figure 3.27B has homologous hits directly where the gene is located which may also be a result of duplications with little translocation.

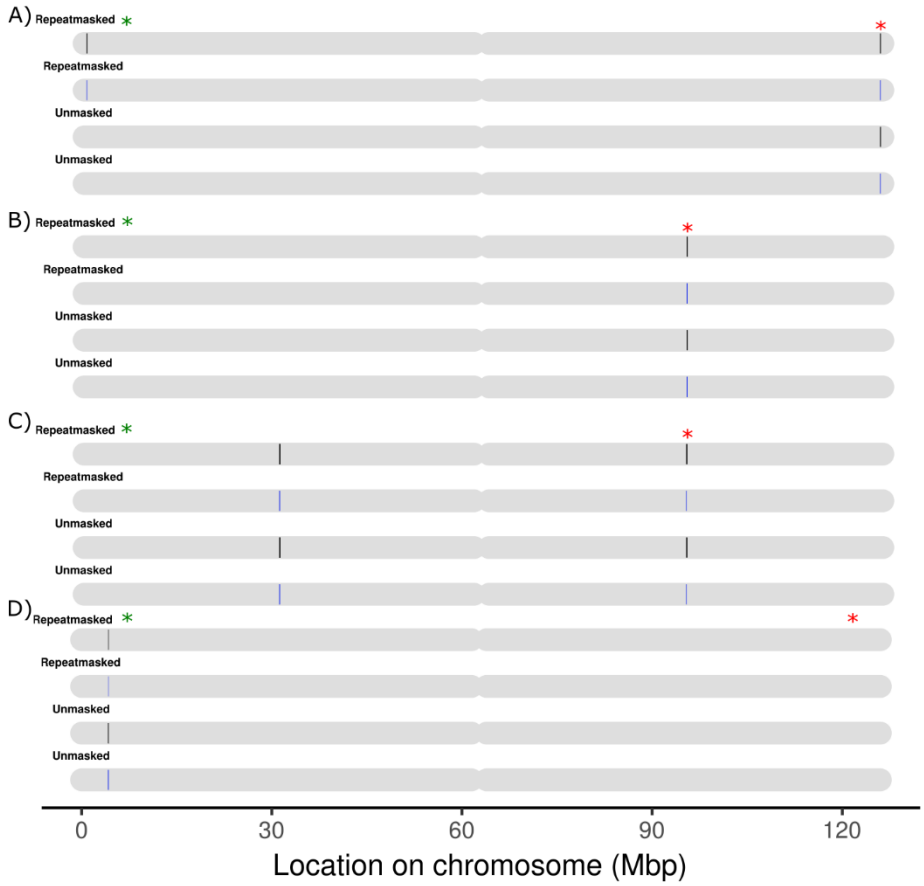

Figure 3.27 Origin of the LASTZ hits in the pig X chromosome overlapping genes. Hits in black are 99% identity and hits in blue are 95% identity. The red asterisk shows the location of the gene to which these hits aligned. The PAR ranges from 0 to the PAR boundary marked with the green asterix. A) Ensembl gene ID ENSSSCG00000034475 B) Ensembl gene ID ENSSSCG00000048704 C) Ensembl gene ID ENSSSCG00000051484 D) Ensembl gene ID ENSSSCG00000040153

The LASTZ hits stacked within regions of the genes are interspersed on the chromosome further suggesting similarities to repetitive elements

The second observed pattern from the segment data showed numerous stacked hits within regions of the gene. The distribution of these hits showed interspersion throughout the chromosome, suggesting their nature to be similar to transposable elements. The nature of many

repetitive elements, such as retrotransposons, is to move within genomes causing them to be interspersed throughout the chromosome yet being most concentrated at the centromere (Cordaux & Batzer, 2009; Wright et al., 2017). This also supports the findings where the greatest density of the interspersed hits were found at the centromere (Figure 3.28) and many of the genes with these stacked hits were located near the centromere (Figure 3.14B).

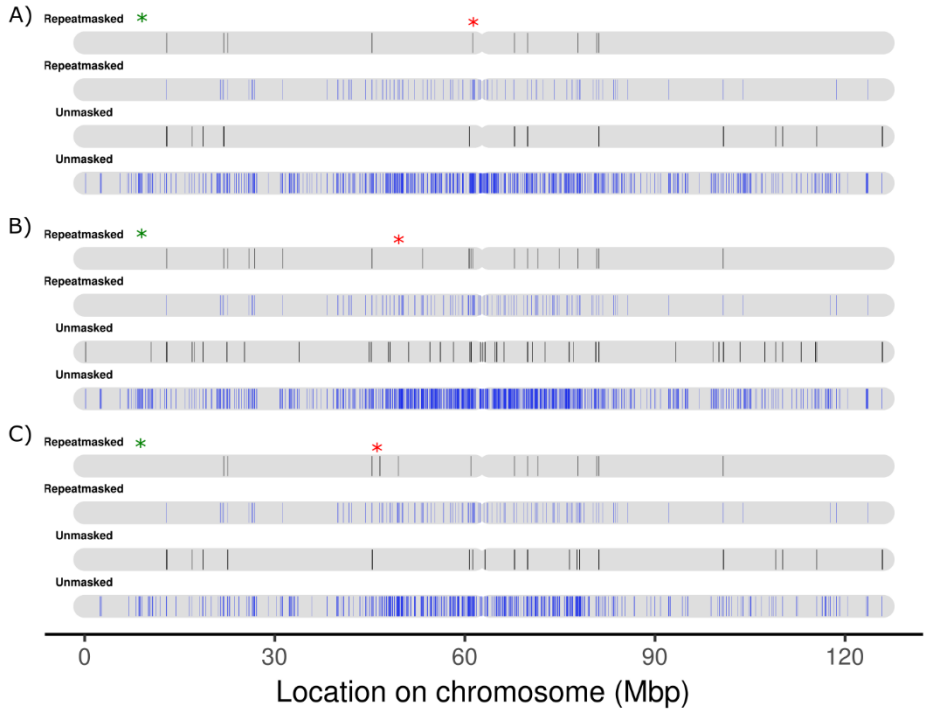

Figure 3.28 Original distribution of LASTZ hits in the pig X chromosome overlapping genes. Hits in black are 99% identity and hits in blue are 95% identity. The red asterisk shows the location of the gene to which these hits aligned. The PAR ranges from 0 to the PAR boundary marked with the green asterix. A) hits homologous to the gene PBDC1 B) hits homologous to the gene SPIN3 C) hits homologous to the gene SMC1A

Genes partially covered by the hits show properties similar to both duplicated genes and repetitive elements

There have been distinct behavioural patterns thus far with the segment groups (genes entirely covered & with embedded stacked alignments) and the distribution of their hits. The alignments partially covering the genes however have distributions similar to both previous patterns. If there has been no change in the size and number of the hits overlapping the gene after

repeat-masking had been applied (Figure 3.15); then the hit distribution shows discrete locations of origin (Figure 3.29A) similar to the hits overlapping entire genes (Figure 3.27).

Where numerous hits partially covered the gene (Figure 3.16) and repeat-masking reduced the hit count/size; there were distribution patterns resembling both the discrete hit locations (Figure 3.27) and the interspersed hits (Figure 3.28). This meaning the unmasked data had some hits spanning the chromosome arms; however, upon repeat-masking the hits showed two discrete locations of origin at 30Mbp and 95Mbp (Figure 3.29B). The reduction in the hits (size and count) is likely the result of repetitive elements being removed. These factors considered, many of the genes with hits partially covering them may be the result of duplicated genes and in some cases also containing embedded repetitive elements.

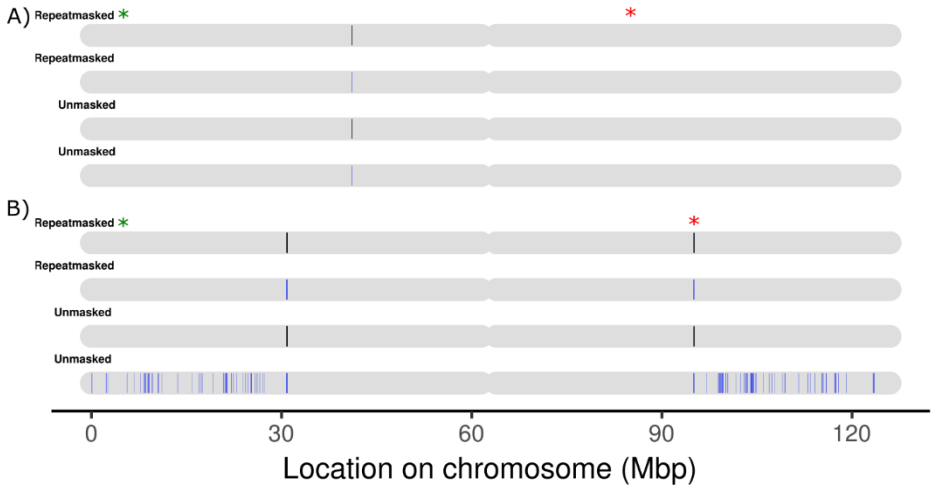

Figure 3.29 Original distribution of LASTZ hits in the pig X chromosome overlapping the genes. Hits in black are 99% identity and hits in blue are 95% identity. The red asterisk shows the location of the gene to which these hits aligned. The PAR ranges from 0 to the PAR boundary marked with the green asterix. A) hits homologous to the gene with the Ensembl ID ENSSSCG00000012517 B) hits homologous to the gene with the Ensembl ID ENSSSCG00000044950

Genes with only unmasked hits overlapping them either have interspersed hits – likely repetitive elements or discrete hit locations – possibly duplicated genes

There have been genes found to have unmasked LASTZ hits and no repeat-masked hits overlapping them. This is not unexpected as repeat-masking removes known repetitive elements which are detected in the alignment. This is clearly the case where there are stacked alignment hits

within genes for only the unmasked data; this stacking of numerous hits suggests the presence of a highly repetitive element (Figure 3.18 and Figure 3.19). This is further confirmed by the distribution of these hits throughout the chromosome (Figure 3.30A & B). Following the known distribution patterns of repetitive elements to be interspersed in the chromosome, the hits support the suggestion these were known repetitive elements removed with repeat-masking.

A proportion of the hits from this grouping were singular hits covering either the entire gene or a part of the gene (Figure 3.20 and Figure 3.21). The hit covering the entire gene may be a duplication where the gene is carried as part of a larger alignment block. The smaller hit however covers a small proportion of the gene and as the entire gene is formed of an exon this may be a region of interest that has been duplicated. The removal of these sequences after repeat-masking however suggests they may either be known repetitive elements or as they are not showing the typical behaviours of highly repetitive elements (Figure 3.30C & D), they may also be low complexity regions of DNA. Further investigation is required to determine their nature.

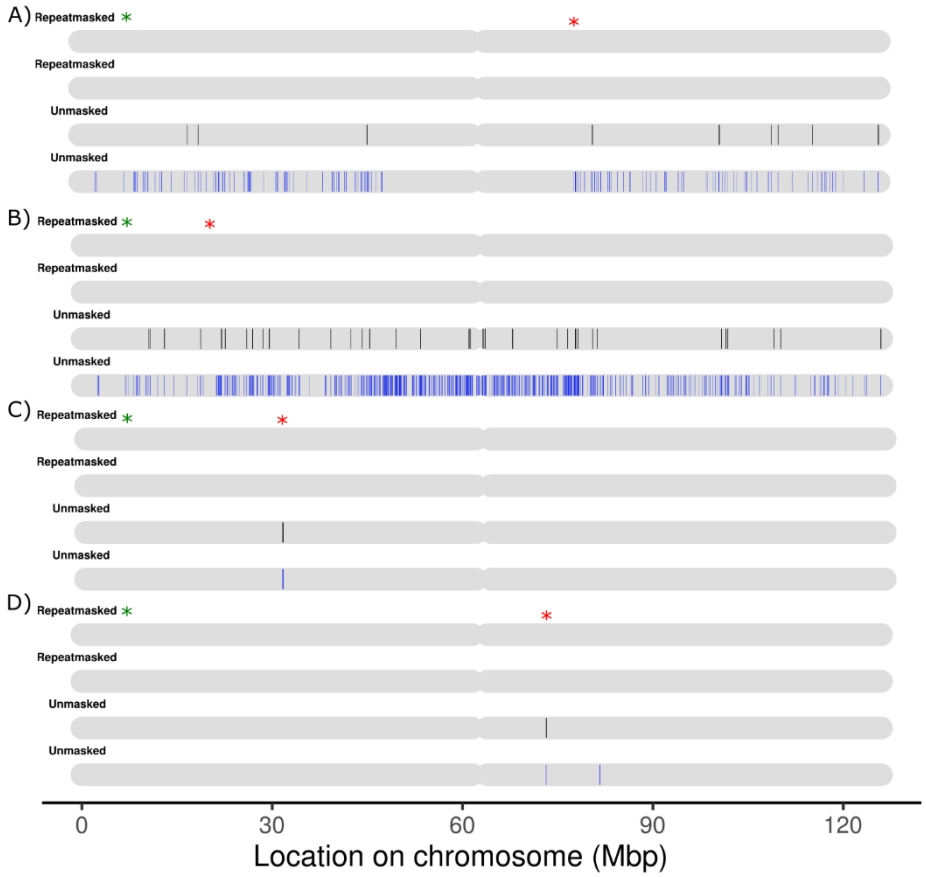

Figure 3.30 Original distribution of LASTZ hits in the pig X chromosome overlapping the genes. Hits in black are 99% identity and hits in blue are 95% identity. The red asterisk shows the location of the gene to which these hits aligned. The PAR ranges from 0 to the PAR boundary marked with the green asterix. A) hits homologous to the gene with the Ensembl ID ENSSSCG00000042737 B) hits homologous to the gene with the Ensembl ID ENSSSCG00000012175 C) hits homologous to the gene with the Ensembl ID ENSSSCG00000036038 D) hits homologous to the gene with the Ensembl ID ENSSSCG00000020695

Staggered alignment hits are likely duplicated genes broken due to incomplete chromosome assemblies

A notable pattern in the segment graphs showed hits staggered throughout the gene (Figure 3.22 Figure 3.23). Not many genes followed this pattern, and these genes were located at 45Mbp with their hits originating primarily from around 47Mbp (Figure 3.31). There were multiple staggered hits in the segment graphs yet only 1-2 locations for the hit origin within their ideograms (Figure 3.31). Having only two locations for multiple hits could suggest these hits are fragments of these regions on the chromosome, whether separated during the alignment or due to incomplete contig assemblies. This could explain why the entire gene was covered by the fragments and how the hits appear to resemble the patterns seen with the hits overlapping entire genes.

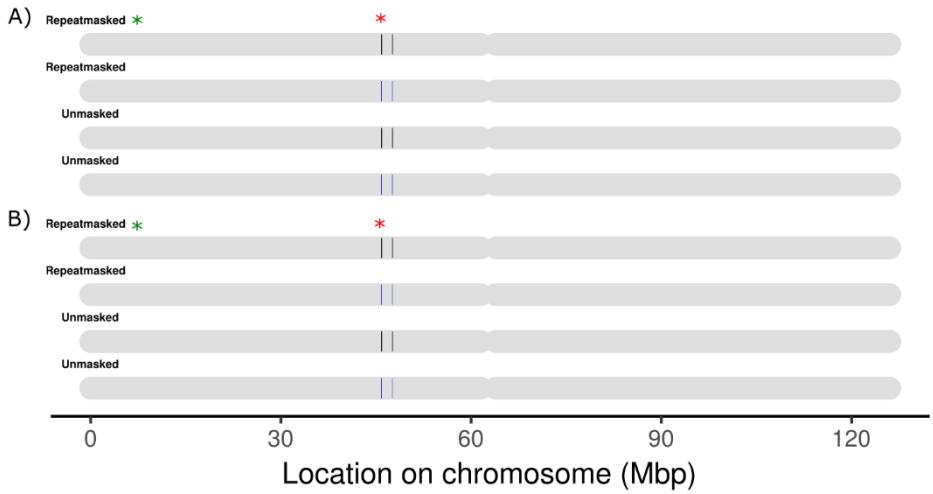

Figure 3.31 Original distribution of LASTZ hits in the pig X chromosome overlapping the genes. Hits in black are 99% identity and hits in blue are 95% identity. The red asterisk shows the location of the gene to which these hits aligned. The PAR ranges from 0 to the PAR boundary marked with the green asterix. A) hits homologous to the gene with the Ensembl ID ENSSSCG00000048218 B) hits homologous to the gene with the Ensembl ID ENSSSCG00000042077

The locations of the genes and hits for both A and B in Figure 3.31 being the same, suggest these hits are reciprocals. This meaning the hits homologous to gene A may have originated from gene B and vice versa. As mentioned before, the distribution of these hits resembles that of the hits which had overlapped entire genes; and may show either a fragmented assembly with the halves of the genes identified in different regions, or a duplication formed through the fragmented alignments.

Genes with overlapping repeat-masked hits and no unmasked hits mostly appear to be fragmented alignments from discrete locations on the chromosome

As mentioned previously, there were some alignments which showed overlaps between 99% identity repeat-masked data and some genes but no overlaps for the 99% identity unmasked data. This is likely a result of the unmasked hits not reaching the required scoring threshold set in the alignment process being dropped to reduce the computational demands. The majority of the hits that followed this pattern were either staggered throughout the gene with partial coverage, such as in Figure 3.24, or had varied numbers of very small hits. The distribution of these hits however was mostly shown as discrete locations in the chromosome (Figure 3.32), suggesting they

could be duplications of genes or genomic elements. Further investigation is required to determine their nature.

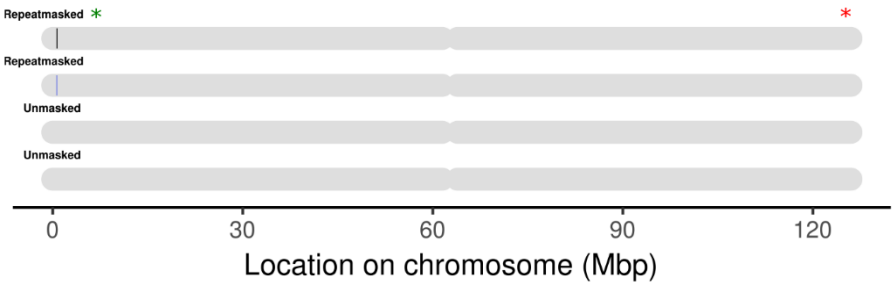

Figure 3.32 Original distribution of LASTZ hits in the pig X chromosome overlapping the gene. Hits in black are 99% identity and hits in blue are 95% identity. The red asterisk shows the location of the gene to which these hits aligned. The PAR ranges from 0 to the PAR boundary marked with the green asterix. A) hits homologous to the gene with the Ensembl ID ENSSSCG00000031635

VII. Highest frequency BLAST subject sequences suggest numerous genes show similarity to LASTZ hits along with repetitive elements such as retrotransposon L1

Given the above, we have identified distinct patterns of LASTZ hits, of which some are overlapping known genes. These patterns showed altered coverage of each gene as seen in the segment graphs (Figure 3.8 Figure 3.24). The original distributions of the hits homologous to these genes also showed altered patterns corresponding to the overlap patterns e.g., the hits stacked within the genes were interspersed throughout the chromosome whereas the hits overlapping entire genes were found in discrete locations. This distribution of the hits testified further that the nature of the hits is likely repetitive elements and duplicated genes. However, this is speculation derived from the known behaviours of repetitive elements.

The nature of this study has been to identify ampliconic genes, and to determine whether any of the replicated genes are ampliconic, we need to determine their nature. Through NCBI BLAST we were able to determine genes and genomic features which showed homology to our LASTZ hits. The nature of ampliconic genes suggests that the most frequently occurring subject sequences returning from the BLAST search would be significant, therefore, the frequency of all

the subject sequences was calculated using their accession numbers for both the unmasked (Figure 3.33A) and repeat-masked (Figure 3.33B) datasets and the top results were assessed.

Numerous subject sequences were returned with varied frequencies between the unmasked and repeat-masked datasets; and in some cases, entirely different subject sequences were observed between the two. Both datasets shared their highest frequency subject sequence - MK028166 (Figure 3.33,Table 3.3 and 3.4). Following this, divergences between the unmasked and repeat-masked data can be seen. For example, the next most frequent subject sequence in the unmasked data is the myostatin gene (AY208121, Figure 3.33A,Table 3.3) whereas in the masked sequence it is an uncharacterised locus (Figure 3.33B, Table 3.4). However, the myostatin gene is still found within the top 5 most frequently occurring subject sequences of the repeat-masked data.

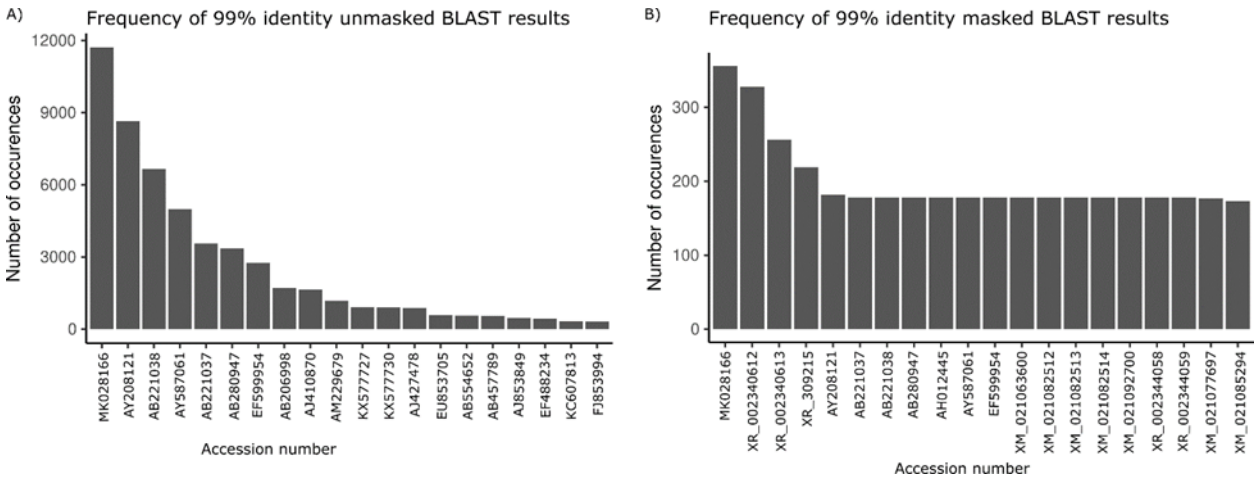

Figure 3.33 Frequency of BLAST subject sequences represented by accession number shown in the X axis A) 99% identity unmasked LASTZ hits and B) 99% identity repeat-masked hits

There were numerous predicted genes and uncharacterised loci found only within the highest frequency subject sequences of the Repeat-masked data, this is also true where only the unmasked dataset has high frequencies of genes such as liver carboxylesterase 1 (PLE-1). These differences in frequencies may have occurred where there were hits in some datasets and not others (i.e., where repeat-masking removed some hits or the unmasked hits were dropped in the scoring threshold of the alignment).

At this stage the potential nature of the homology found within the X chromosome has been determined. However, there are a large number of genes returned from the BLAST data and it is unlikely they are all duplicated genes in the chromosome. The genes detected in the BLAST search probably harbour genomic elements which are homologous to the LASTZ hits - for example, LINE elements that are often found inserted within genes.

Table 3.3 Highest frequency subject sequences from the NCBI BLAST nt search on the hits from the 99% identity unmasked LASTZ data

Subject title	Accession number	Frequency
WIF1 (WIF1) gene, partial sequence; and LEM domain-containing 3 (LEMD3) and methionine sulfoxide reductase B3 isoform X1 (MSRB3) genes	MK028166	11714
myostatin gene	AY208121	8649
cd1b, cd1e, or10t1, or10j1, or10k1, or10r1, or10k2, or10r2 genes for CD1B antigen, CD1E antigen, putative olfactory receptor 10T1, putative olfactory receptor 10J1, putative olfactory receptor 10K1, putative olfactory receptor 10R1, putative olfactory receptor 10K2, putative olfactory receptor 10R2	AB221038	6659
ataxia-telangiectasia mutated protein gene, exons 1 through 64 and complete cds	AY587061	4988
cd1d, cd1b, cd1e, or10t1 genes for CD1D antigen, CD1B antigen, CD1E antigen, putative olfactory receptor 10T1	AB221037	3561
DCT, GPC6 genes for dopachrome tautomerase, glypican 6	AB280947	3353
retrotransposon L1, complete sequence	EF599954	2752
ASIP and AHCY genes for agouti signaling protein and S-adenosylhomocysteine hydrolase	AB206998	1706
COX7A1 gene, CAPNS1 gene, CKAP1 gene, POLR2I gene and CLIPR-59 gene (partial)	AJ410870	1649
IFNAR1 gene, IFNGR2 gene and TMEM50B gene	AM229679	1179
liver carboxylesterase 1 (PLE-1) gene	KX577727	908
liver carboxylesterase G2 (PLE-G2) gene	KX577730	901
ASIP gene for agouti signalling protein and AHCY gene for S-adenosylhomocysteine hydrolase	AJ427478	878
acetyl-coenzyme A carboxylase beta (ACACB) gene, alternatively spliced	EU853705	578
VRTN gene for vertnin and around region, strain: LWD	AB554652	561
DNA, TRDV gene segments for T cell receptor delta-chain	AB457789	555
PRSS11 gene (partial), DMBT1 gene, AWN gene, AQN-1 gene, PSP-II gene, PSP-I gene, AQN-3 gene and C10ORF120 gene	AJ853849	458
adenylate kinase 3-like 1 (AK3L1) gene	EF488234	439
domesticus integrin beta 5 (ITGB5) gene	KC607813	330
FTo gene	FJ853994	314

Table 3.4 Highest frequency subject sequences from the NCBI BLAST nt search on the hits from the 99% identity hard Repeat-masked LASTZ data

Subject title	Accession number	Frequency
WIF1 (WIF1) gene, partial sequence; and LEM domain-containing 3 (LEMD3) and methionine sulfoxide reductase B3 isoform X1 (MSRB3) genes	MK028166	356
PREDICTED: uncharacterized LOC110257707 (LOC110257707), transcript variant X1, ncRNA	XR_002340612	328
PREDICTED: uncharacterized LOC110257707 (LOC110257707), transcript variant X2, ncRNA	XR_002340613	256
PREDICTED: uncharacterized LOC102166571 (LOC102166571), ncRNA	XR_309215	219
myostatin gene	AY208121	182
cd1d, cd1b, cd1e, or10t1 genes for CD1D antigen, CD1B antigen, CD1E antigen, putative olfactory receptor 10T1	AB221037	178
cd1b, cd1e, or10t1, or10j1, or10k1, or10r1, or10k2, or10r2 genes for CD1B antigen, CD1E antigen, putative olfactory receptor 10T1, putative olfactory receptor 10J1, putative olfactory receptor 10K1, putative olfactory receptor 10R1, putative olfactory receptor 10K2, putative olfactory receptor 10R2	AB221038	178
DCT, GPC6 genes for dopachrome tautomerase, glypican 6	AB280947	178
locus 1q2.4 SBAB 130A12 PERV-A LTRs and flanking genomic regions	AH012445	178
ataxia-telangiectasia mutated protein gene, exons 1 through 64 and	AY587061	178
retrotransposon L1	EF599954	178
PREDICTED: olfactory receptor 2A1/2A42-like (LOC100525417), mRNA	XM_021063600	178
PREDICTED: leucine carboxyl methyltransferase 1-like (LOC100625764), transcript variant X1, mRNA	XM_021082512	178
PREDICTED: leucine carboxyl methyltransferase 1-like (LOC100625764), transcript variant X2, mRNA	XM_021082513	178
PREDICTED: leucine carboxyl methyltransferase 1-like (LOC100625764), transcript variant X3, mRNA	XM_021082514	178
PREDICTED: katanin catalytic subunit A1 like 2 (KATNAL2), transcript variant X3, mRNA	XM_021092700	178
PREDICTED: uncharacterized LOC102167067 (LOC102167067), transcript variant X1, ncRNA	XR_002344058	178
PREDICTED: uncharacterized LOC102167067 (LOC102167067), transcript variant X2, ncRNA	XR_002344059	178
PREDICTED: uncharacterized LOC100624414 (LOC100624414), mRNA	XM_021077697	177
PREDICTED: uncharacterized LOC100517025 (LOC100517025), transcript variant X9, mRNA	XM_021085294	173

VIII. Further investigation of most frequent BLAST subject sequences confirms LINE elements to be the source of much of the homology

The BLAST tables of results provided information pertaining the subject sequences homologous to regions of the LASTZ alignments. The next stage of the analysis was to determine the likely cause of the homology between the subjects and the LASTZ hits. With many different genes having been returned the true nature of the hits needed to be isolated, such as whether the homology shown was aligned to an entire gene or genomic element. The DNA sequence within the query start and stop positions (the region of the LASTZ hit homologous to the subject sequence) was extracted and aligned to the subject sequences to determine the significance of the similarity.

The majority of the most frequently found BLAST subject sequences are embedded retrotransposon L1 sequences

Most of the highest frequency BLAST subject sequences showed homology to many of the LASTZ hits. These similarities were found to be within discrete regions of the subject sequences, often with small alignments stacked in exact locations. This was true both in the repeat-masked and unmasked datasets; the main difference being when repeat-masking was applied the number and size of the homologous regions decreased. This pattern would suggest many of the subjects from the initial BLAST analysis of the LASTZ hits have originated from a similar source. A few examples of this pattern were shown where the 99% identity LASTZ hits were aligned to the pig accession regions for: WIF1 (Figure 3.34A), myostatin (Figure 3.35A), and retrotransposon L1 (Figure 3.36A). These subject sequences also had the 99% repeat-masked LASTZ sequences aligned to them (Figure 3.34B, Figure 3.35B, Figure 3.36).

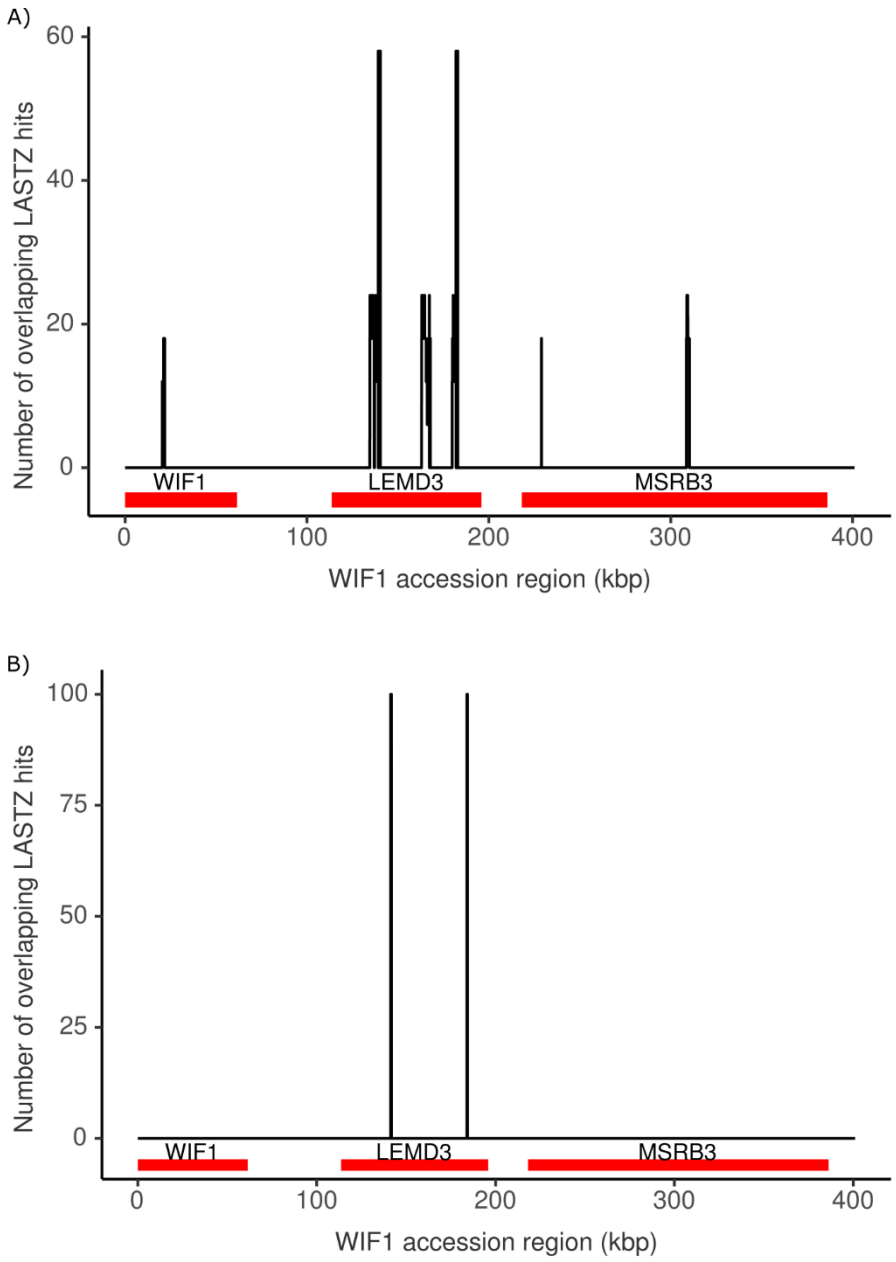

Figure 3.34 LASTZ hits homologous to the most frequently occurring NCBI BLAST subject sequences aligned to the accession sequence for the pig WIF1, LEMD3, MSRB3 region. Within the accession region there were 3 genes present (WIF1, LEMD3, and MSRB3) shown in red, the positions of the genes were extracted from NCBI GenBank. The accession region was broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits. A) 99% identity unmasked LASTZ hits and B) 99% identity repeat-masked LASTZ hits

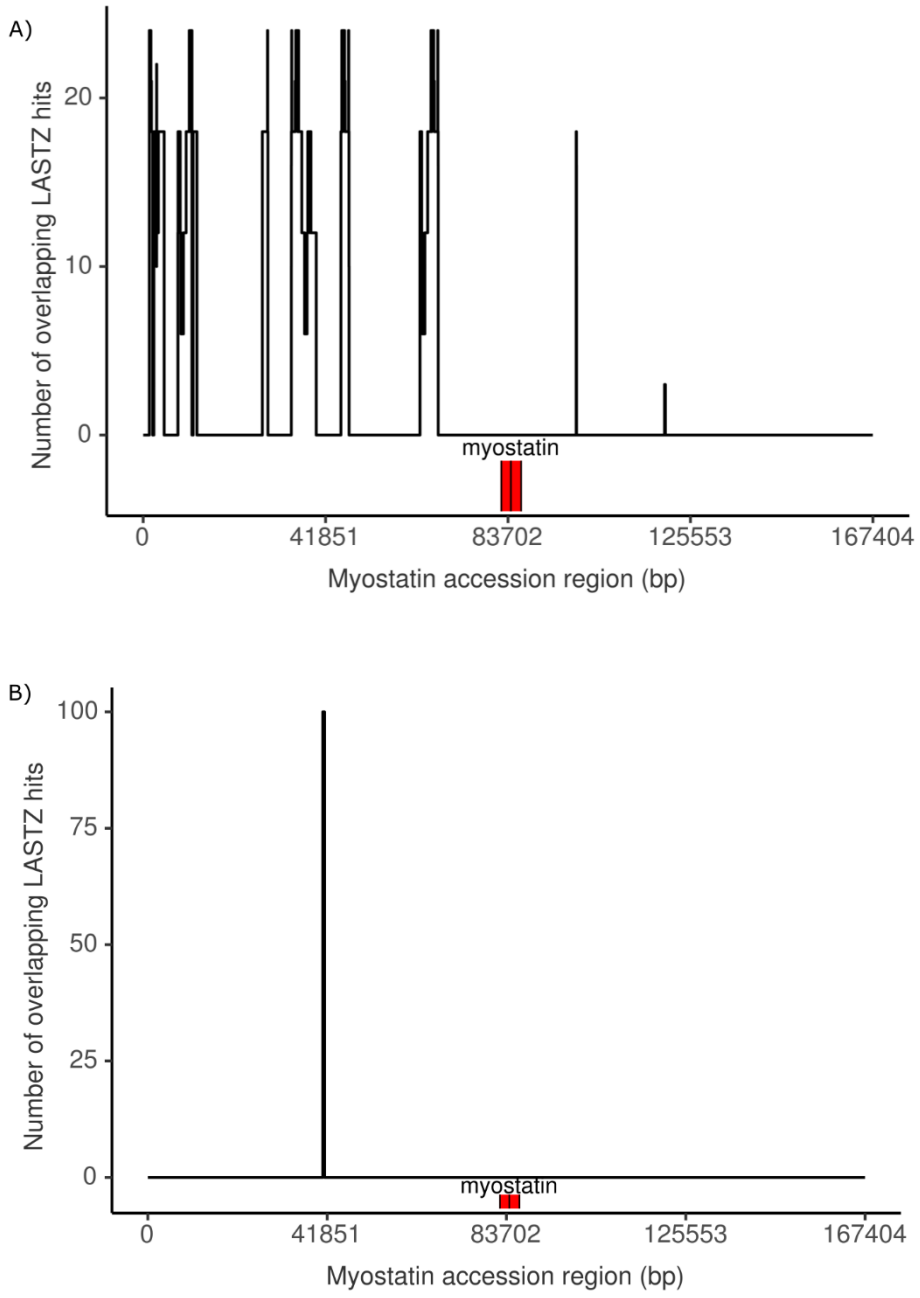

Figure 3.35 LASTZ hits homologous to the most frequently occurring NCBI BLAST subject sequences aligned to the accession sequence for the pig myostatin region. Within the accession region there was only the myostatin gene shown in red, with the exons drawn in black upon the red, the positions of the genes and exons were extracted from NCBI GenBank. The accession region was broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits. A) 99% identity unmasked LASTZ hits and B) 99% identity repeat-masked LASTZ hits

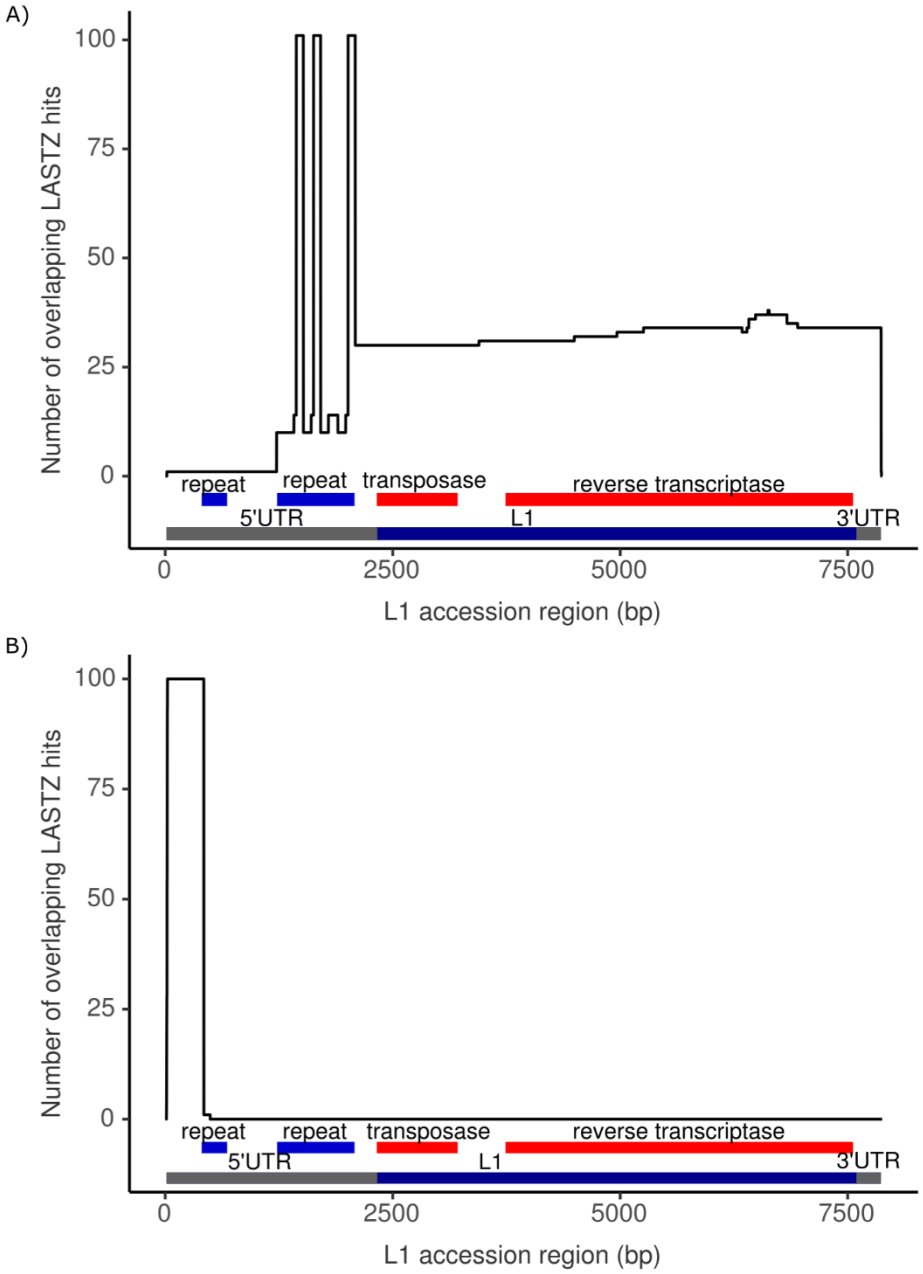

Figure 3.36 LASTZ hits homologous to the most frequently occurring NCBI BLAST subject sequences aligned to the accession sequence for the pig L1 region. Using the annotations from NCBI GenBank, the repeat regions were annotated in blue, the transposase and endonuclease/reverse transcriptase genes were annotated in red, the retrotransposon L1 was annotated in dark blue with the 5'UTR and 3'UTR overlapping in grey. The accession region was broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits. A) 99% identity unmasked LASTZ hits and B) 99% identity repeat-masked LASTZ hits.

Notably when aligning the LASTZ sequences to the WIF1 and myostatin accession regions there were small regions in which the majority of the hits showed similarity (Figure 3.34, Figure 3.35). The alignments to L1 however showed the unmasked hits to cover the majority of the region (Figure 3.36A) with exception of the start at the 5'UTR region, however the repeat-masked hits

exclusively covered this region (Figure 3.36B). This suggests many of the hits to be the result of L1 sequences as the homology appears more significant than the WIF1 and myostatin coverage. It is also likely the fragments shown in the repeat-masked dataset are the result of unannotated L1 fragments which may have been dropped in previous alignments; as mentioned previously, the full unmasked pig DNA sequence provides a high computational demand for the alignment software.

Further suggesting the nature of these hits to be L1 sequences is the comparison to the WIF1 and myostatin alignments where the homology is often within intronic genes (Figure 3.34) or lying outside of the genes entirely (Figure 3.35). However not all of the LASTZ hits showed homology to the L1 sequences, suggesting their nature to be of a different origin.

The 'outlier' genes in the frequency data showed little homology to the commonly found subject sequences and no homology to L1 sequences

The frequency data from the BLAST results for the repeat-masked and unmasked datasets (Figure 3.33, Table 3.3 and 3.4) showed many similarities but also some differences. The main differences being the *IFNAR1* and *COX7A1* genes in the highest frequency subjects for the unmasked data; and a selection of uncharacterised loci in the highest frequency subjects for the repeat-masked data. The LASTZ hits homologous to these genes/loci were aligned to the most frequent BLAST subject sequences found between the datasets to determine if there was any homology shared.

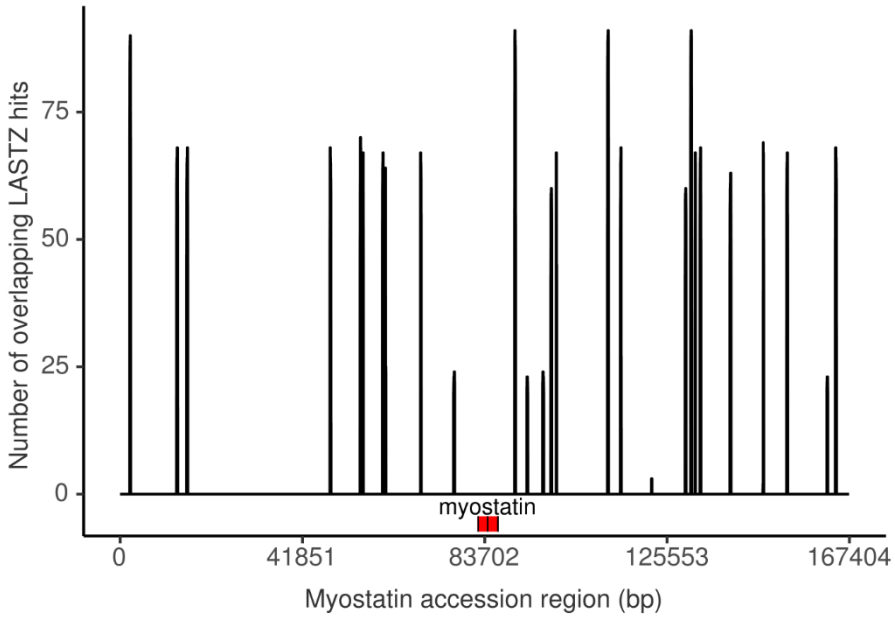

Figure 3.37 Unmasked 99% identity alignment hits from the subject sequences IFNAR1 and COX7A1 aligned to the accession sequence for the pig myostatin gene. Within the accession region there was only the myostatin gene shown in red, with the exons drawn in black upon the red, the positions of the genes and exons were extracted from NCBI GenBank. The accession region was broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits.

The unmasked LASTZ hits homologous to *IFNAR1* and *COX7A1* shared some homology to the common BLAST subject sequences such as the myostatin accession region (Figure 3.37), suggesting there may be embedded similarities. The assumption was this was again the result of L1 sequences however when the hits were aligned to the accession region there were no similarities detected. Similar to Figure 3.35, the alignments between the hits and subject sequences were often outside of the genes or within intronic regions. This may suggest the hits to be low complexity DNA which would be removed from the repeat-masked dataset.

The repeat-masked LASTZ hits homologous to uncharacterised loci, however, did not show any homology to the common BLAST subject sequences including L1. These are therefore likely to be regions of homology with no relation to the repetitive elements as detected within the other subject sequences.

This analysis of the LASTZ hits and their homologous BLAST subject sequences was a generalised overview suggesting the nature of many of the hits to be repetitive elements. However,

this generalised picture does not connect which subject sequences coincide with the patterns of the hits as seen in the segment graphs and ideograms. Further alignments between the individual LASTZ hits and BLAST subject sequences can identify what genes/genomic elements are responsible for the homology.

BLAST results for individual LASTZ hits confirm the majority of the hits to be either repetitive elements or duplicated genes

The LASTZ hits showing self-homology in the pig X chromosome have primarily been found to be small hits stacked within the gene or large hits covering the entire gene; these hits have then been found distributed throughout the chromosome or in discrete locations (respectively). The general NCBI BLAST results of the LASTZ hits showed the hits to be primarily L1 sequences however some hits showed no homology to these L1s, suggesting they are of a different nature. L1 sequences are to be expected where some of the segment patterns appeared to be highly repetitive (e.g., Figure 3.11), however, some of the graphs suggested there to be duplicated genes (e.g., Figure 3.8). Determining the nature of the hits showing these patterns requires the individual LASTZ hits to be BLASTed and aligned to the genes.

LASTZ hits covering the entire gene and partially covering the gene show duplicated regions of the chromosome containing varied genomic elements

The LASTZ hits partially or entirely covering the genes were located discretely in the chromosome with a varied number of alignment hits. These patterns along with the BLAST results from these alignment hits suggest there has been a duplication of the gene or segment carrying the gene. Firstly, the hits overlapping the entire gene at 125Mbp (Figure 3.8) showed homology to predicted gene: histone H2A-Bbd type 2/3 like. This gene was found in the *Sus scrofa* genome and many other species.

The LASTZ alignments have been shown to cover the entire exon for the Histone H2A-Bbd type 2/3 like in *Sus scrofa* (Figure 3.38A); and has partial coverage for the orthologues of the histone (e.g., Figure 3.38B). The duplication is larger than the gene showing a segmental duplication carrying the gene and its exon; suggesting a significant duplication has occurred. Patterns similar to this, with entire gene duplications, have been seen with similar BLAST results covering the majority or entirety of the gene. Another example of a duplication where the LASTZ hit covered the entire X chromosome gene, ENSSSCG00000040153. The BLAST results showed homology with various species of heat shock transcription factor X-linked like from various species, particularly the ferret (Figure 3.9). The ferret HSFX gene has been partially overlapped suggesting a small region of similarity.

Some of the LASTZ hits which overlapped entire genes were homologous to uncharacterised loci from numerous species. These hits covered the majority of the *Sus scrofa* uncharacterised loci LOC110257707 (Figure 3.40A) with regions of similarity in other species (e.g., Figure 3.40B). This locus was homologous with the genes in Figure 3.9, Figure 3.10, and Figure 3.16. Interestingly, these genes are found in the 95Mbp region which showed to be likely gene duplications/inversions in Figure 3.3. The distributions of these hits are primarily around 30Mbp and 95Mbp suggesting some overlap (Figure 3.27 B & C, and Figure 3.29B). This locus was investigated further using shoot.bio, the phylogenetic search engine (Emms & Kelly, 2021), there were no phylogenetic branches detected.

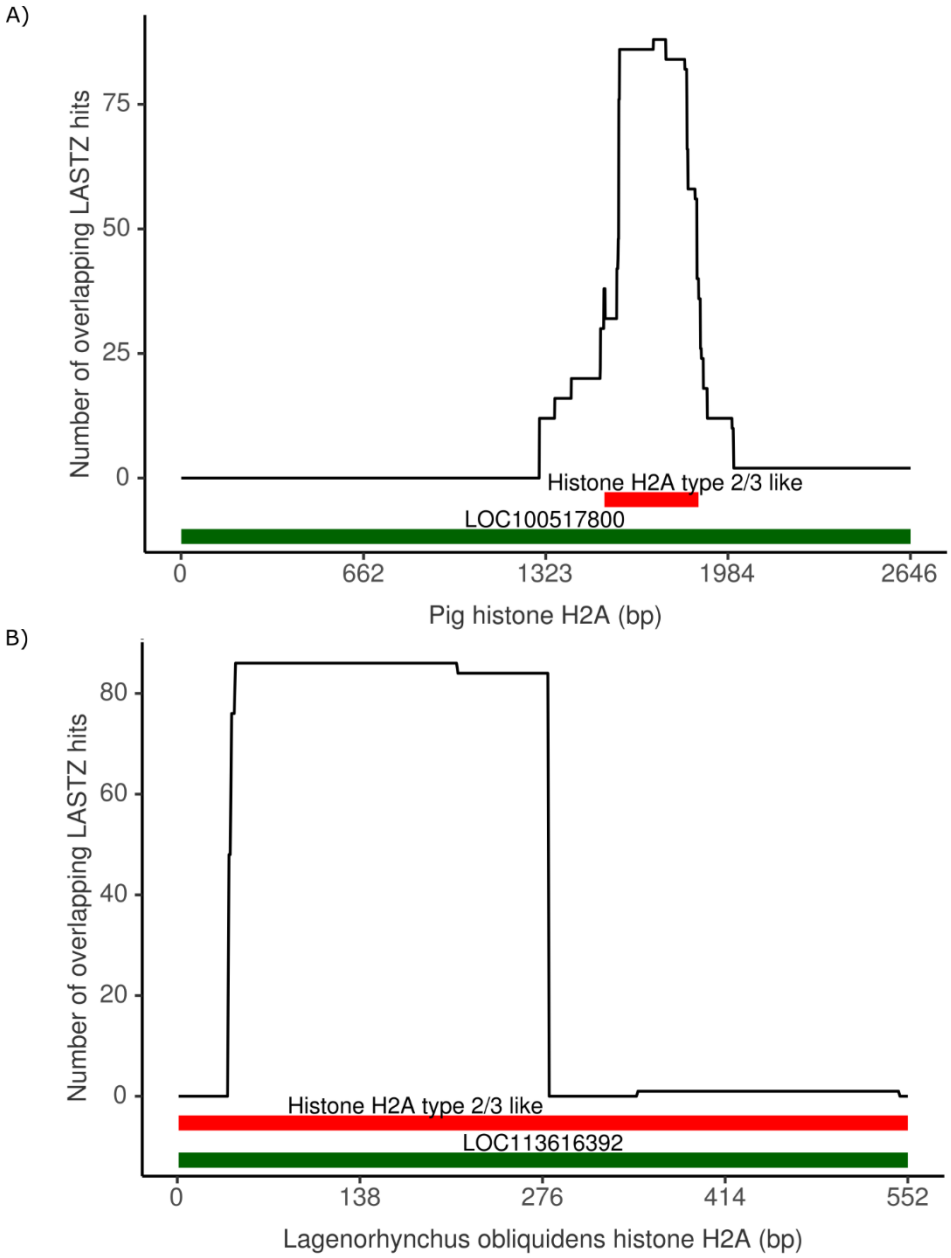

Figure 3.38 LASTZ alignment hits homologous to the gene with the Ensembl ID ENSSSCG00000034475 showed homology to the A) Sus scrofa Histone H2A-Bbd type 2/3 like accession region and B) Lagenorhynchus obliquidens Histone H2A-Bbd type 2/3 like accession region. Using the annotations from NCBI GenBank, the accession region locus was annotated in green, the histone H2A type 2/3 like gene was annotated in red. The accession region was broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

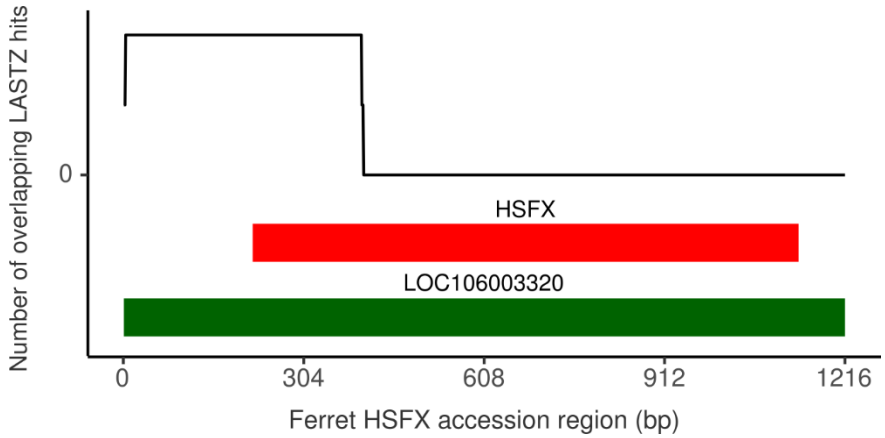

Figure 3.39 LASTZ alignment hits homologous to the gene with the Ensembl ID ENSSSCG00000040153 showed homology to the Mustela putorius furo heat shock transcription factor X-linked accession region. Using the annotations from NCBI GenBank, the accession region locus was annotated in green, the heat shock transcription factor X-linked gene was annotated in red.. The accession region was broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

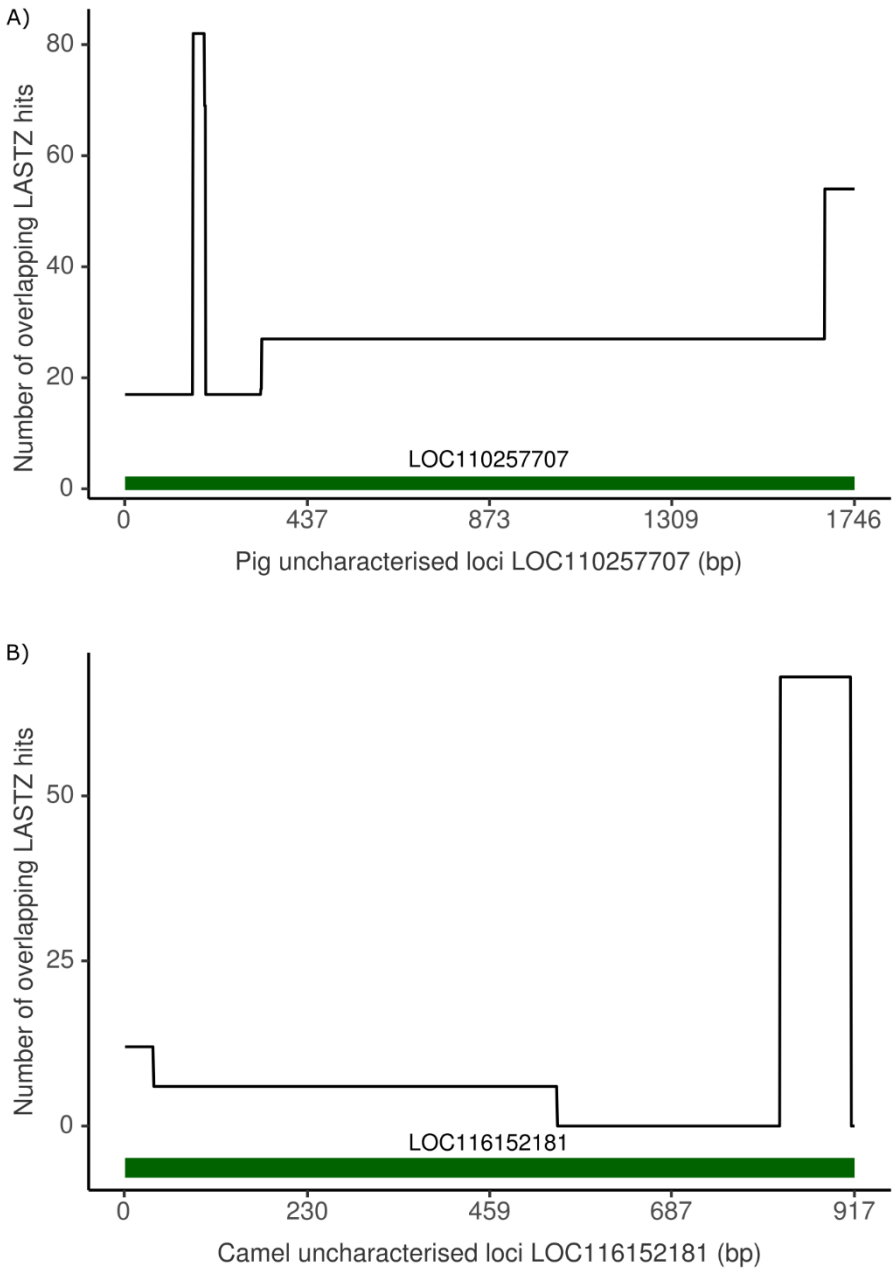

Figure 3.40 LASTZ alignment hits homologous to the genes with the Ensembl ID ENSSSCG00000048704, ENSSSCG00000051484, and ENSSSCG00000044950 showed homology to the subject sequence of uncharacterised loci A) Sus Scrofa LOC110257707 B) Camelus dromearius LOC116152181. Using the annotations from NCBI GenBank, the accession region locus was annotated in green. The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

LASTZ hits stacked in specific regions within the X chromosome genes are fragments of L1 sequences embedded within the genes

Next, investigating the LASTZ hits stacked within regions of the X chromosome genes, which appeared to be the result of highly repetitive elements, the BLAST results were very similar to those commonly found in the general BLAST analysis (Table 3.3, Table 3.4, Figure 3.33). As

seen in Figure 3.34, Figure 3.35, Figure 3.36 many of these genes were homologous due to embedded L1 sequences. The hits stacked within the genes showed similar alignment patterns; particularly where the repeat-masked hits showed to be fragments of the of the L1 sequence located within the 5'UTR (Figure 3.41).The genes which the L1 sequences appear to be embedded include PBDC1 (Figure 3.11), SPIN3 (Figure 3.12), and SMC1A (Figure 3.13).

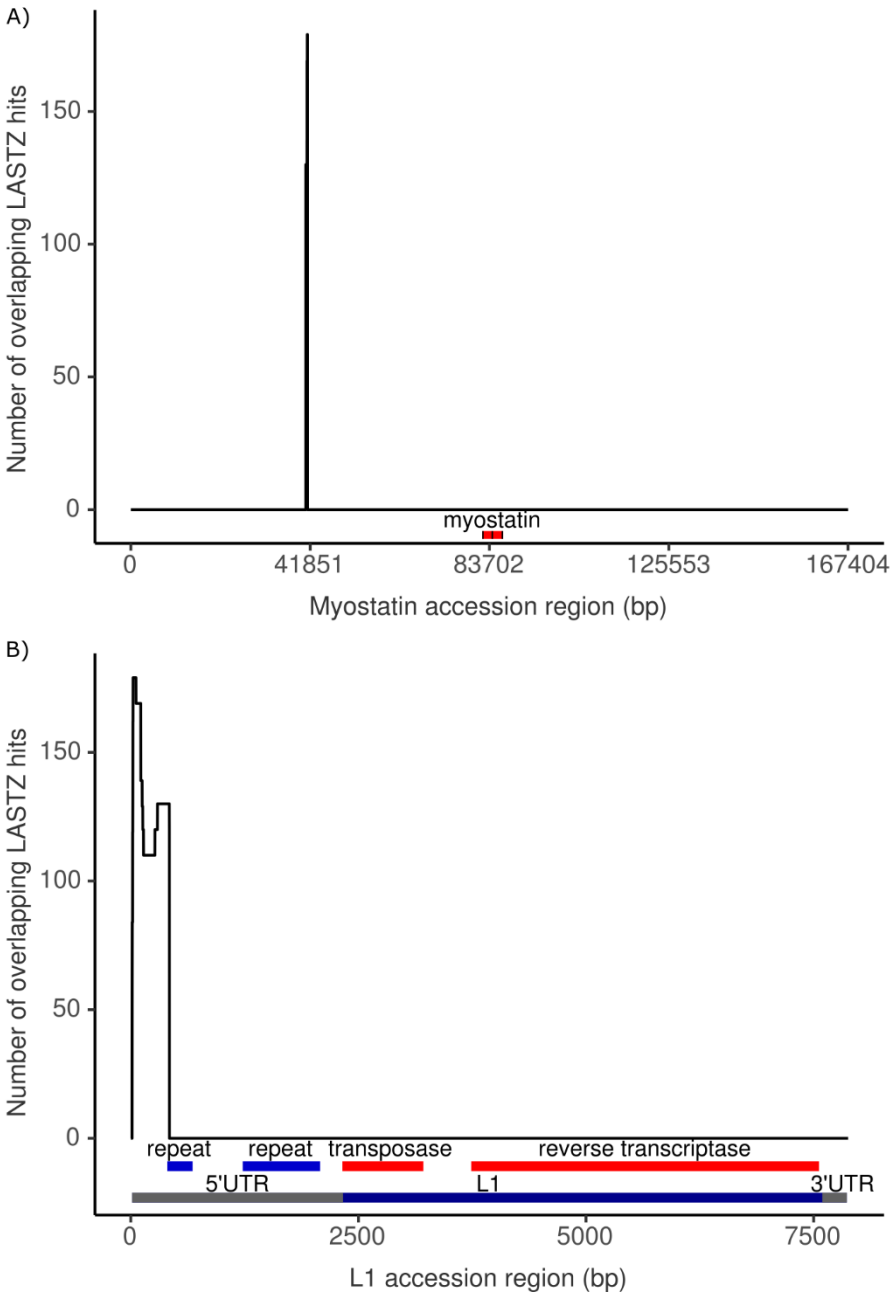

Figure 3.41 LASTZ alignment hits homologous to the genes PBDC1, SPIN3, and SMC1A showed homology to the subject sequence of A) myostatin gene B) retrotransposon L1. Within the myostatin accession region there was only the myostatin gene shown in red, with the exons drawn in black upon the red, the positions of the genes and exons were extracted from NCBI GenBank. Using the annotations from NCBI GenBank the L1 accession region was annotated, the repeat regions in blue, the transposase and

endonuclease/reverse transcriptase genes in red, the retrotransposon L1 in dark blue with the 5'UTR and 3'UTR overlapping in grey. The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

The hits overlapping the genes within only the unmasked datasets are confirmed to be known repetitive elements

As previously mentioned, some genes had only unmasked hits overlapping them; this was suggested to be the result of previously annotated repetitive elements, and this was confirmed through the BLAST investigation of these hits. The returned subject sequences for the unmasked hits stacked within the gene showed homology to many familiar genes such as the WIF1 accession region, the myostatin gene, and retrotransposon L1. As previously seen in Figure 3.34, Figure 3.35, Figure 3.36, and Figure 3.41; the hits covered the majority of the L1 sequence and only small discrete regions within the gene accession, often outside of the gene itself (Figure 3.42). This suggests these hits to be the result of L1 sequences.

Not all of the unmasked hits showed homology to these L1 sequences however, there were some singular large hits such as in Figure 3.20 covering the entire gene, and upon further investigation these hits appeared to be genes such as DCT and CBX6. This homology is interspersed throughout the accession regions for these genes (Figure 3.43). The small size of these homologous hits, their interspersion in the gene, having found them within the introns and outside the genes, and their removal with repeat-masking; suggests the homology between the hits and these genes is the result of low complexity DNA.

A proportion of the 99% identity unmasked hits, such as in Figure 3.21, have shown to be homologous to known repetitive elements other than retrotransposons. In this case, the subject sequences returned were variants of the porcine endogenous retroviruses which when aligned to the hits revealed the LASTZ alignments had detected the long terminal repeats in each PERV (Figure 3.44).

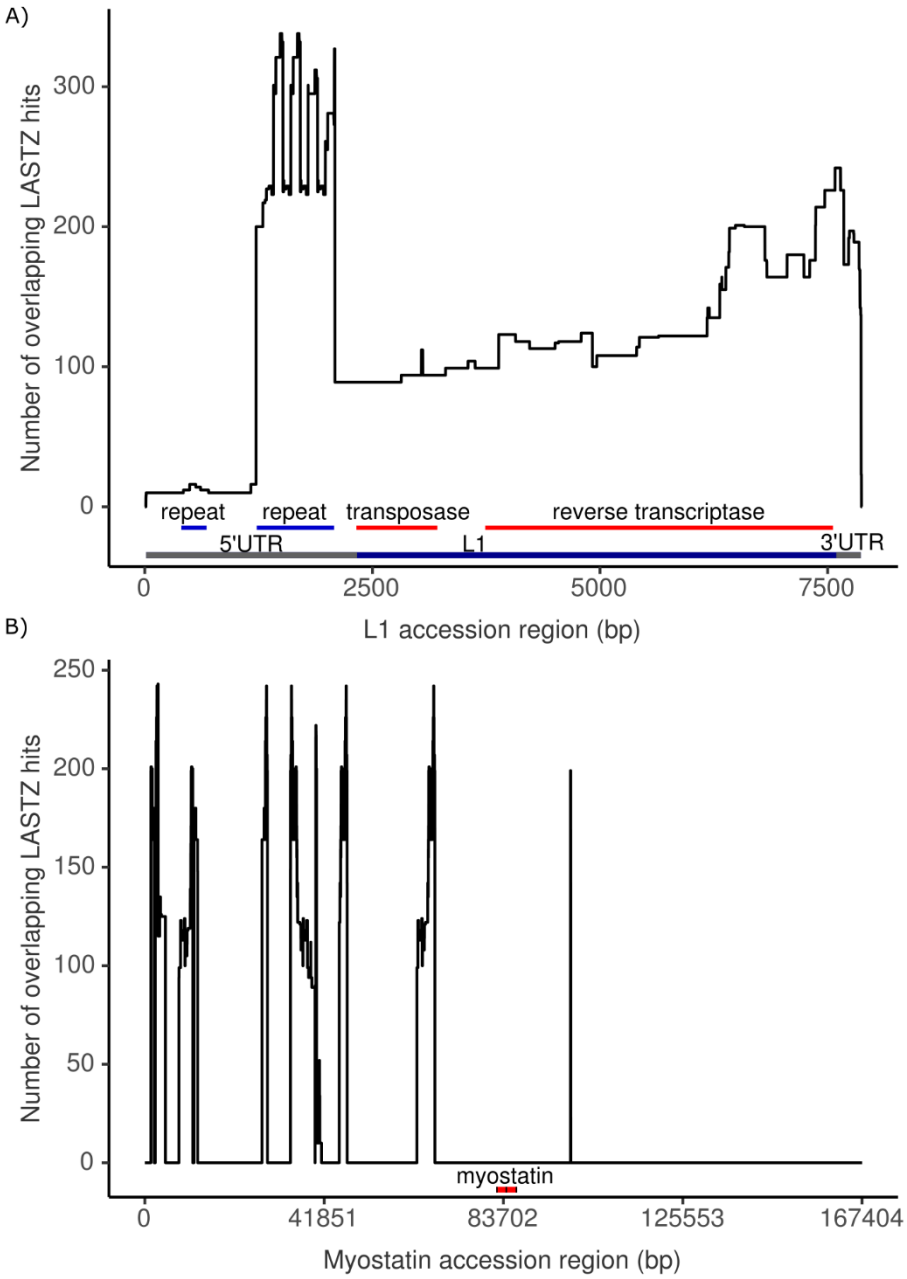

Figure 3.42 99% identity unmasked LASTZ alignment hits homologous to the gene with the Ensembl ID ENSSSCG00000042737 showed homology to the subject sequences of A) retrotransposon L1 B) myostatin gene. Within the myostatin accession region there was only the myostatin gene shown in red, with the exons drawn in black upon the red, the positions of the genes and exons were extracted from NCBI GenBank. Using the annotations from NCBI GenBank the L1 accession region was annotated, the repeat regions in blue, the transposase and endonuclease/reverse transcriptase genes in red, the retrotransposon L1 in dark blue with the 5'UTR and 3'UTR overlapping in grey. The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

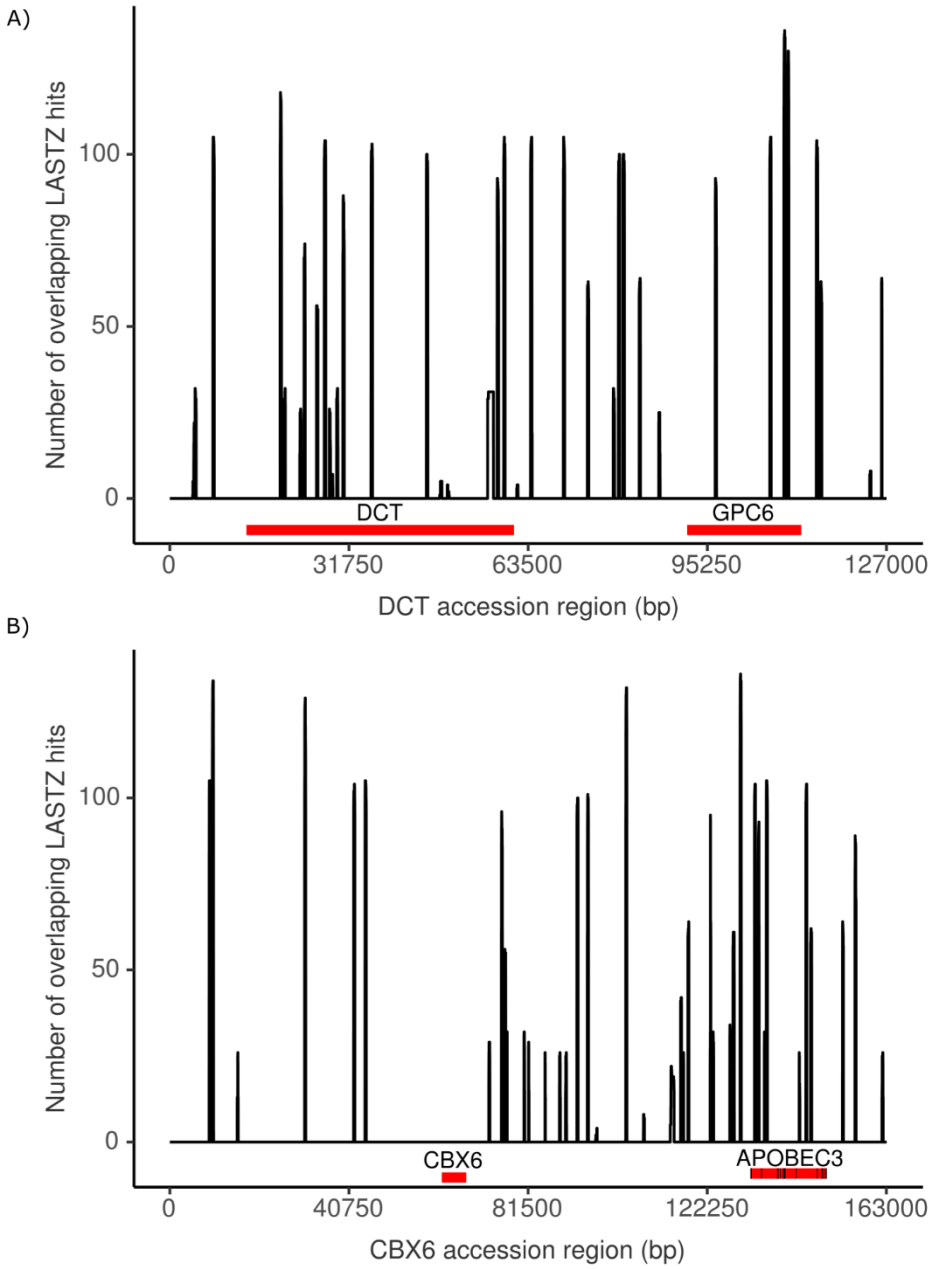

Figure 3.43 99% identity unmasked LASTZ alignment hits homologous to the gene with the Ensembl ID ENSSSCG00000036038 showed homology to the subject sequence of A) DCT B) CBX6. Within the DCT accession region two genes were present (DCT and GPC6) shown in red. Within the CBX6 accession region two genes were present (CBX6 and APOBEC3) shown in red. The positions of the genes were extracted from NCBI GenBank The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

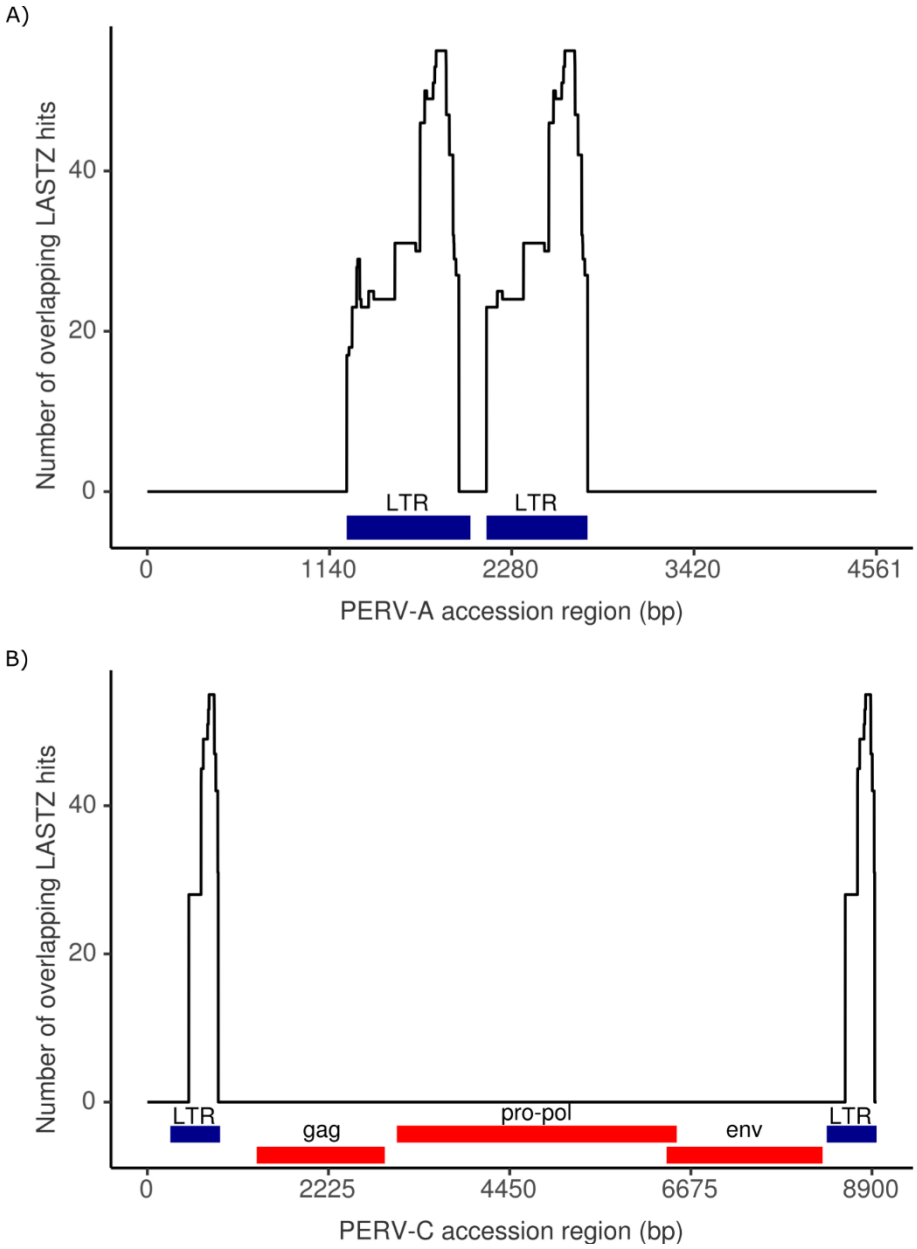

Figure 3.44 99% identity unmasked LASTZ alignment hits homologous to the gene with the Ensembl ID ENSSSCG00000020695 showed homology to the subject sequence of A) locus 1q2.4 SBAB 130A12 PERV-A LTRs and flanking genomic regions B) pig isolate PERV-C endogenous virus Porcine endogenous retrovirus. The PERV-A locus is annotated with two LTRs in dark blue, and the PERV-C region has two LTRs in dark blue and 3 genes in red (gag,pro-pol, and env). The positions of the genes and exons were extracted from NCBI GenBank. The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

LASTZ hits staggered throughout the X chromosome gene show similarities to hits covering entire genes suggesting they are fragments of duplicated genes

The hits staggered throughout the genes (Figure 3.22 and Figure 3.23) behaved similarly to the hits overlapping entire genes with regard to their distribution; this also rang true with the BLAST results showing homology to an uncharacterised locus. The hits from both genes

were located at the 47Mbp region with the same BLAST subject sequences; suggesting these may be the reciprocals of one another where there has been a duplication of this region of the chromosome. The uncharacterised locus was investigated on shoot.bio (Emms & Kelly, 2021) and showed no phylogenetic branches, although the BLAST results showed numerous orthologues present. The LASTZ hits were homologous to a large proportion of the uncharacterised locus and therefore is highly likely this locus has been duplicated (Figure 3.45).

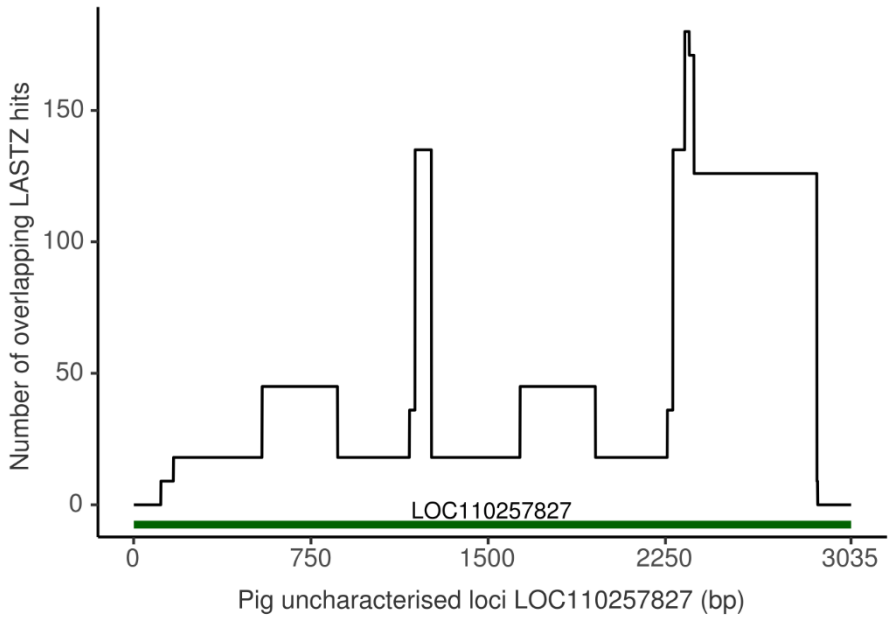

Figure 3.45 LASTZ alignment hits homologous to the genes with the Ensembl ID ENSSSCG00000042077 and ENSSSCG00000048218 showed homology to the subject sequence of the uncharacterised locus LOC110257027. Using the annotations from NCBI GenBank, the accession region locus was annotated in green. The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits

Repeat-masked hits partially covering the genes with no unmasked reciprocals are similar to some genomic regions however the smaller hits appeared to be low complexity DNA

As seen in Figure 3.24, some genes only had overlapping hits from the repeat-masked dataset; much of this is likely due to the hspthresh scoring parameter that had been defined (see 2.IV). These hits fell into two main patterns; partially covering the gene (e.g., Figure 3.24) or smaller hits varied in size and number. Investigating the small hits revealed them to overlap introns

and showed no (or very low) similarity when run through BLAST suggesting they are low complexity regions of DNA.

The hits partially covering the genes (Figure 3.24) however appear to be duplications of genomic regions; as seen in the alignments (Figure 3.46) where the hits showed homology to various orthologues of the coagulation factor VIII associated 1 (F8A1) and its intron 22 protein. Many of the alignments cover the intron 22 protein as a small region of the gene suggesting this region has been duplicated: likely suggesting its presence in a pig variant of the F8A1 gene.

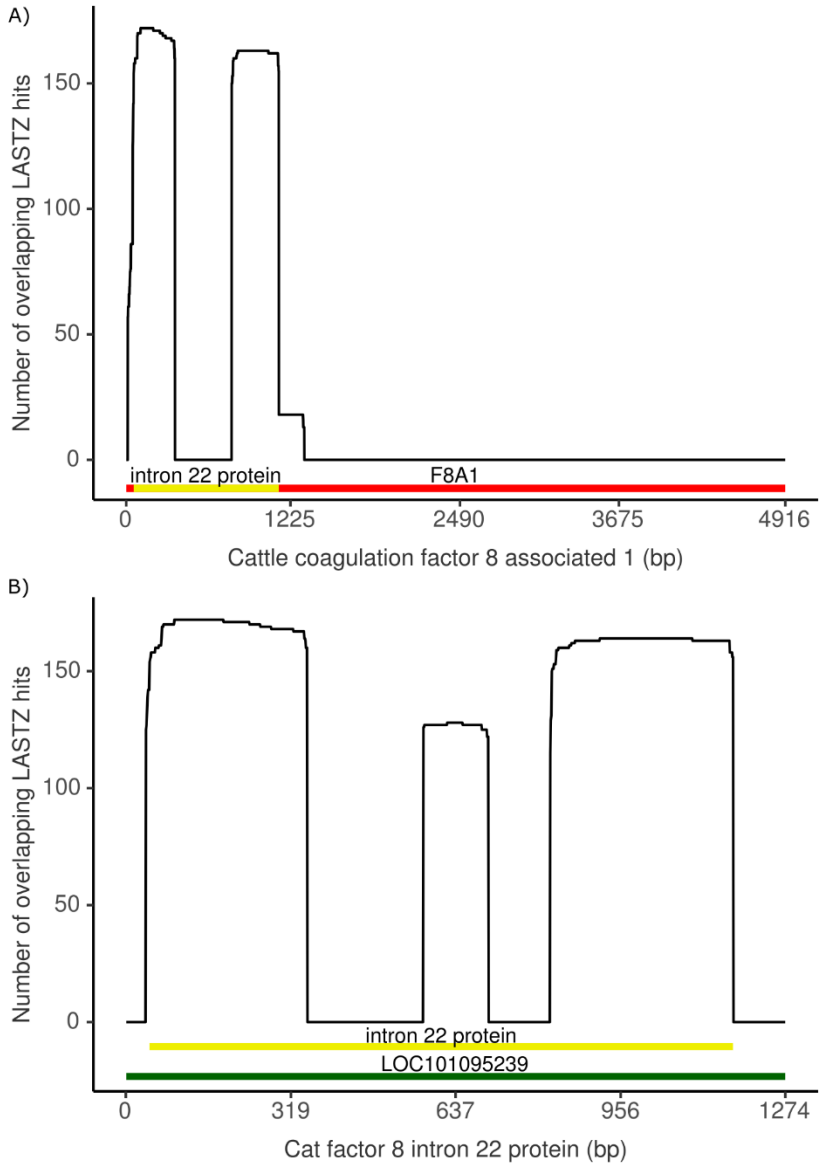

Figure 3.46 LASTZ alignment hits homologous to the gene with the Ensembl ID ENSSSCG00000031635 showed homology to the subject sequence of A) Bos indicus x Bos taurus coagulation factor VIII associated 1 (F8A1) B) Felis catus factor VIII intron 22

protein. Using the annotations from NCBI GenBank, the Felis catus accession region locus was annotated in green, and the intron 22 protein in yellow. The Bos indicus x Bos taurus accession region F8A1 gene was annotated in red with the intron 22 protein overlapped in yellow. The accession regions were broken into windows of 1bp with the number of LASTZ hits overlapping each window calculated and plot as a line graph showing the extent of the homology between the accession region and the LASTZ hits.

Insertions of L1 sequences into *Sus scrofa* genes is a recent occurrence in the evolution of the pig X chromosome

Many of the accession regions assessed showed to have L1 sequences within or around their genes suggesting transposition has occurred at some stage in the genome. Determining when this may have occurred required aligning the L1 sequence to these genes in other species. Here we looked at humans, mice, and bulls which all have good quality genome assemblies (see 1.VII). The accession regions which showed to have L1 sequences in the BLAST search (e.g., *WIF1* and myostatin), showed no alignment similarities between the *Sus scrofa* L1 sequence and any of the gene orthologues. This may suggest that the L1 insertion occurred more recently in the ancestry of the pig genome and therefore are likely novel insertions. There were also no L1 insertions detected in the orthologues of the X chromosome genes which the LASTZ hits had overlapped. This includes *PBDC1*, *SPIN3*, and *SMC1A*. These genes are homologues between the species and pigs however the insertions of the L1 fragments have likely occurred recently within the pig lineage.

IX. Summary

The results of this study have shown no significant palindromes under 1000bp in length within the Sscrofa11.1 assembly; however, larger duplications and possible inversions were identified at 45Mbp and 95Mbp. Further investigation showed the pig X chromosome to be dense with self-homology; much of which was the result of highly repetitive elements such as retrotransposon L1's. These L1 insertions are specific to the pig lineage and appear to show recent transposition as a sign of 'recent' evolution of the pig genome. Other than the L1 insertions, the self-homology has also been the result of duplicated genes in the pig X chromosome. Many of these genes showed homology to other pig genes and their orthologues such as the histone H2A-Bbd type 2/3 like, and

the coagulation factor VIII associated 1 (F8A1) and its intron 22 protein. However, there were also many uncharacterised loci which appear to be duplicated, found within the regions identified with large duplications and possible inversions (45Mbp and 95Mbp). At this stage, no evidence has been found for ampliconic genes in the pig X chromosome; however, identification of repetitive elements will prove useful in updating the current annotation of the chromosome, and the duplicated genes, in particular, the uncharacterised loci provide interesting candidates for further investigation.

4. Discussion

This thesis aimed to identify any ampliconic genes present in the pig X chromosome. No evidence was found for amplicons, however, the extent of the repetitive landscape of the pig X chromosome assembly Sscrofa11.1 was determined. The repetitive content identified fell into two distinct categories: 1) duplicated/multicopy genes and 2) highly repetitive elements such as retrotransposons. These results and previous literature can shed a light on these replications and begin to determine why they may have occurred; and what they may mean in the context of genome evolution.

I. IUPAC showed no significant palindromes in pig X chromosome however large duplications found in preliminary LASTZ alignment

DNA palindromes increase susceptibility to genome instability where response mechanisms to double stand breaks (DSBs) lead to chromosomal rearrangements, such as translocations, deletions or gene amplifications with various consequences; particularly longer inverted repeats with high sequence identity and shorter spacing between palindrome arms (Miklenic & Svetec, 2021). Only 12 palindromes were detected in the pig Sscrofa11.1 X chromosome assembly found between 55bp and 119bp in length (Table 3.1). These palindromes appear functionally insignificant due to their small size and locations between genes or within gene introns. Functional palindromes as seen in humans and mice are much larger with large inverted repeats around 11kb in length with 99% sequence identity as observed in physical maps (Small et al., 1997).

Further studies using modernised computational methods determined 26 large IRs in the human X chromosome ranging from 8-140kb in length all with over 99% sequence identity (Jackson et al., 2020; Warburton et al., 2004). Twelve of these are shared in primates such as chimpanzees and rhesus macaques as highly conserved on the X chromosome having a common origin up to 25 million years ago (Jackson et al., 2020). Mueller et al., (2008) found mice have 17

palindromes (above 90% identity, >8kb in length) within 22 ampliconic regions: further showing the presence of long IRs within mammalian sex chromosomes. This suggests there are few significant inverted repeats found in the X chromosome of pigs – at least in the current assembly. The IUPAC software was developed to be less computationally demanding than EMBOSS in detecting inverted repeats (Alamro et al., 2021); however, this was not the case in this study where increasing the default values exponentially increased run-time. Reassessing the pig X chromosome inverted repeat content with a different software may reveal larger palindromes throughout the chromosome as potential regions to identify ampliconic genes.

X chromosome duplications and inversions were seen in the preliminary LASTZ analysis exceeding 10-15kb in length (Figure 3.3); these are of a more similar size to the X ampliconic sequences found in humans and mice (Mueller et al., 2013). There has been no evidence for ampliconic genes within inverted repeats in the pig X chromosome, however, there have been some duplications identified. Obtaining accurate reference sequences can be difficult, leading to complications in determining inverted repeat content; this may influence the lack of evidence for amplicons, in contrast to humans and mice, in the incomplete Sscrofa11.1 X chromosome assembly. However, the regions of the pig X chromosome which have yet to be assembled are small and are unlikely to harbour amplicons, although not impossible.

II. Arrangement of self-alignment hits through LASTZ support findings in cytogenetic studies

The pig X chromosome is dense with self-homology particularly around 20-30Mbp, 35-50Mbp, 60-70Mbp, 75-80Mbp, and 95-110Mbp (Figure 3.6); this correlates with numerous cytogenetic studies of the pig genome. Substantial constitutive heterochromatin was observed through C-banding on the X chromosome; this was found within the interstitial arms, surrounding the centromere, and near the telomeres (Adega et al., 2005). High densities of constitutive heterochromatin were identified at the centromere (Hansen-Melander & Melander, 1974);

coinciding with the intensity of the hits located at the centromere and the regions of density found within the chromosome arms in this study. There have also been associations made between G-banding pattens (a type of chromosome staining to identify heterochromatin) and chromosomal locations of mammalian transposable elements such as the retrovirus (long interspersed nuclear elements) LINEs (Wichman et al., 1993). The GTG-banded pig karyotype as depicted by Frönicke & Wienberg, (2001) coincides well with many of the regions dense with self-homology as shown in the LASTZ alignment; further confirming the LINE coverage in the pig X chromosome.

Array-CGH studies (detecting copy number variations in a genome) also found regions of likely repetitive content within the pig X chromosome. These regions were at similar points: 20-30Mbp, 40-50Mbp, and ~90Mbp (Skinner et al., 2013). There were also copy number variation regions noted throughout the X chromosome suggesting chromosomal rearrangements and duplications (Wang et al., 2014).

The density of repetitive content within the centromeric region of the chromosome is consistent between these studies; strong heterochromatin content has been associated with regions of suppressed recombination, and the centromere is known as a region devoid of homologous recombination (Talbert & Henikoff, 2010; Wright et al., 2017). Eukaryotic centromeres have been seen to be rich with repetitive sequences; with heterochromatin flanking chromatin structures which are involved in kinetochore assembly (Zafar et al., 2017). This would explain the consistent findings of C-band density within the centromere of the pig X chromosome. Regions, such as the pericentromere, associated with transposable elements such as LINEs have also been found with dense constitutive heterochromatin (Slotkin & Martienssen, 2007).

The centromere and pericentromere were not alone in the detection of constitutive heterochromatin; both the LASTZ alignments and the banding patterns of other studies detected

telomere densities. It has been suggested that short tandem repeats are found at telomere ends and the sub-telomere harbours constitutive heterochromatin similar to pericentromeric heterochromatin containing both full length and fragmented transposable elements (Slotkin & Martienssen, 2007).

III. LINEs prominent throughout the pig X chromosome may alter gene expression and provide insights into genome evolution

Much of the self-homology detected in the pig X chromosome was found to be the result of unannotated transposable elements; these findings were consistent with regions of constitutive heterochromatin isolated by cytogenetics studies. Much of the homology between the transposable elements showed to be embedded within gene introns (Figure 3.11, Figure 3.12, &Figure 3.13) and originating from various regions spread across the chromosome (Figure 3.28).

Transposable elements are found dispersed throughout chromosomes mostly within introns and between genes

It's agreed across studies that L1 density is vast in the X chromosome; correlating with the results from the LASTZ analysis on the pig X. Transposable elements (TEs) such as L1s are often large genomic regions which have been replicated and dispersed throughout a genome and often have high sequence identity (Lander et al., 2001). Mammalian TEs are formed of four main groups; long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), LTR retrotransposons and DNA transposons (Lander et al., 2001). Retrotransposons such as LINEs 'copy and paste' themselves as a mechanism to move through a genome (Bourque et al., 2018). TEs are often considered to be 'selfish' or 'parasitic' as their high levels of replication cause them to accumulate in the genome often posing negative consequences to the host fitness; however most TEs are inactive due to deletions and mutations although, some full length active TEs remain (Slotkin & Martienssen, 2007). Our study found fragments of unannotated L1's in the repeat-

masked dataset and full-length L1s in the unmasked dataset of the pig X chromosome above 99% identity

The localisations of these L1s in mammals is not random. Firstly there is often a preference of TEs for insertion into host exons; deleterious insertions may be selected against, however, non-deleterious insertions may become fixed between genes and within introns - interspersed throughout the chromosome (Bourque et al., 2018). This is what we found: L1s embedded within gene introns and interspersed throughout the chromosome.

The accumulation of TEs at centromeres and pericentromeric regions (see 4.II) may be the result of an insertion bias for regions of high heterochromatin content; or where TEs may be selected against due to their effect on genes and accumulation in regions of reduced recombination reduces the likelihood of spreading via genetic drift; or the hill-Robertson interference (see 1.0) may lead to accumulations and fixations of TEs (Wright et al., 2017). These suggestions also apply to the accumulation of L1s in the X chromosome, which often does not recombine with the Y chromosome; allowing for these accumulations to occur throughout the chromosome.

Fragments of L1s not masked by RepeatMasker show further work is needed to annotate species-specific transposable elements

Repeat-masked DNA sequences obtained from Ensembl used RepeatMasker to identify repetitive elements and low complexity DNA. Why were fragments of pig L1 sequences not masked? The TE sequence information used by RepeatMasker comes from Repbase Genetic Information Research Institute (Smit et al., 1999), a database of eukaryotic TEs and repeats where consensus sequences are reconstructed for families of TEs. The most recently available documentation of the database contents revealed the constructed database was formed from a total of 134 species which included only 28 mammalian species - as of 2015 (Bao et al., 2015).

RepeatMasker also stores genomic datasets including *Sus scrofa*; documenting 827 ancestral families of TEs and 112 species specific families (*Pig [Sus Scrofa] Genomic Dataset*, n.d.). However, it is not stated as to the individual elements identified. Highly conserved TE's are well documented and sufficiently identified, however, where there are lineage specific divergences alignment success decreases (Koning et al., 2011). The pig specific L1 fragments not annotated may be highly diverged leading to unsuccessful alignments; this would require detailed alignments of the pig L1s and other mammalian species to determine.

Computational demands led to alignments being dropped in the LASTZ analysis (see 2.IV); this too may have occurred in the alignment of pig specific L1 sequences leading to fragments escaping annotation. Regardless, these fragments have been identified and can be used to improve the current annotation of the pig X chromosome. Future studies would benefit from ensuring accurate annotations of species-specific transposable elements by improving software databases such as RepeatMasker.

L1 insertion in genes may affect gene expression in the X chromosome

PBDC1, SPIN3, and *SMC1A* were found to have embedded L1 sequences in this study; these protein coding genes located on the X chromosome all have over 200 orthologues in many species according to Ensembl. *SMC1A* and *SPIN3* both play roles in maintenance of structures, however the function of *PBDC1* is not well documented (Kaessmann & Rappold, 2018). *SMC1A* is part of the structural maintenance of chromosomes (SMC) complexes; which play a role in chromosome segregation and maintenance, through highly conserved mechanisms (Yatskevich et al., 2019). The *SPIN3* protein has been associated with mitotic spindle organisation and stability (Cao, 2012).

Dosage compensation mechanisms (outlined in section 1.0) involve one of the female X chromosomes being inactivated regulating the imbalance of gene expression between males and

females (Katsir & Linial, 2019). Interestingly, *PBDC1* has been seen to escape inactivation in mice yet not in humans, the inverse is true of *SMC1A* (Lopes, et al., 2011; Musio, 2020). Escapee genes remain with reduced expression levels in the inactive X chromosome leading to a higher overall expression of these genes in females than males (Berletch et al., 2011; G. Chen et al., 2016; Fang et al., 2019). Disteche (1999) theorised these genes may have a sex-specific function in females; however, this does not seem to be the case where *PBDC1* and *SMC1A* are inactivated in humans or mice respectively. Determining whether there is inactivation of the pig orthologues may be of interest.

Embedded L1's within these genes may suggest susceptibility to inactivation where L1s are suggested to aid X-inactivation (see 1.0). *PBDC1* - inactive in humans has an inserted L1 sequence, whereas the escapee orthologue in mice has no L1 insertion. L1 presence may influence its inactivation. LINE-1 is said to facilitate spreading of *Xist* RNA - an initiator of X inactivation (Pontier & Gribnau, 2011; Tang et al., 2010). Studies have found *Xist* spreading to be reduced in areas of low LINE-1 density, however, the involvement of LINEs in X-inactivation has been vastly debated as there has in fact been LINE enrichment seen around the X inactivation centre in humans but this has not been observed in mice (Pontier & Gribnau, 2011; Tang et al., 2010).

LINEs also control gene expression, cell differentiation, and DNA repair (Ngamphiw et al., 2014). The insertions of L1s to genes have also been suggested to be deleterious where the gene can become inactivated or properties altered; although these effects would likely be selected against (Boissinot et al., 2001). Humans, mice, and pig genomes are abundant with L1s, often with diminished activity in humans and mice due to 5' truncations (Ngamphiw et al., 2014). The LASTZ analysis found fragments of L1s in the repeat-masked dataset homologous to the 5' UTR region (Figure 3.41). The unmasked hits were truncated at the 5' region likely due to dropped alignments (see section 2.III); adding the 5' fragments from the repeat-masked data could reveal

full L1s in the pig X which may be active. However, truncations are not the only cause of L1 inactivation as rearrangements and mutations can also reduce their functionality (Szak et al., 2002).

These functions are often discussed in the context of conserved L1s and it has been suggested more recent L1 insertions may have minor phenotypic consequences (Ngamphiw et al., 2014). No similarity was detected when aligning the *Sus scrofa* L1 to orthologues of *PBDC1*, *SPIN3*, and *SMC1A*, suggesting the insertion in the pig lineage to be recent. Intronic localisation of the L1s would suggest there to be little functional significance. However, our analysis showed fragments from the 5' UTR region which is stressed to be functionally important, and high density of 'recent' L1s suggests recent mobilisation (Ivancevic et al., 2016). Further investigation would be beneficial to determine the active status and functional properties of these L1 sequences.

L1s have evolved to have lineage specific variants influencing evolution of mammalian genomes

Mammalian transposable elements can evolve to become lineage-specific; as in the pig L1 sequence having no homology to other species, this has been observed between mice and human L1 lineages (Rodriguez-Terrones & Torres-Padilla, 2018). Species-specific divergence is a sign of recent mobilisation and can influence the differential evolution of species (Mills et al., 2006). The evolutionary implication of L1 divergence includes genome instability and cancer development/progression in humans; understanding the evolution of these L1's between and within species may provide an insight into genome evolution (Ivancevic et al., 2016).

Determining the active status of the pig specific L1's is important to understand their impact on the chromosome, particularly as they have been found interspersed along the entire pig X chromosome. The presence of these L1's, if active, may influence expression levels of genes with or around inserted L1s. Further investigation such as this may also improve our current understanding of the role of LINEs in X inactivation. Finally, characterising these L1 sequences

and their effect on the pig genome will provide valuable evolutionary insights for pig genomes and other mammalian genomes.

IV. Genes duplicated in the pig X chromosome show evolution and divergence of the pig genome from other mammalian species

The LASTZ analysis showed alignment hits covering entire X chromosome genes within a larger alignment block due to segmental duplications and inversions. These hits were investigated further to determine their nature.

Histone H2a-Bbd is duplicated in the pig X chromosome correlating with other species however the functional role is still unclear

The X chromosome gene ENSSSCG00000034475 located at 125Mbp had two alignment hits overlapping its entirety originating from ~125Mbp and ~8Mbp (Figure 3.8). The Ensembl database identified 94 orthologues of ENSSSCG00000034475 - defined by ratio: 1:many or many:many. This meaning either: within a pair of species there is one gene in one species homologous to many genes in the other species; or there are many homologues in both species. The Ensembl orthologues were found to be histone H2A-Bbd genes correlating with the BLAST results of the overlapping alignment hits showing to the predicted histone H2A-Bbd type 2/3 for pigs and various other species. The human orthologue of ENSSSCG00000034475 is syntenic according to Ensembl however there was no comparison shown for the mouse orthologue.

Histone variants regulate nuclear processes such as transcription, DNA repair, and chromosome segregation (Tolstorukov et al., 2012). H2As in particular are suggested to be evolutionarily young and rapidly evolving - originating on the X chromosome in the last common ancestor of placental mammals (Molaro et al., 2020). Highly expressed in the testis of humans and mice, H2As interact with splicing factors at actively transcribed genes; disruption of the three

H2A-Bbd encoding genes, in mice, manifested chromatin dysfunction and splicing changes (Anuar et al., 2019; Chew et al., 2021). Expression levels of H2A is less understood in pigs.

Mammalian H2A-Bbd variants have 2-3 duplicate copies possibly due to segmental duplications, complementing our findings in the LASTZ analysis. H2A-Bbd's duplicated copy number in humans and mice is required to encode the functional histones has often been shown to involve 3 genes (Chew et al., 2021; Ishibashi et al., 2010). The LASTZ analysis showed two alignment hits overlapping the gene suggesting three total copies, this similarity suggests this gene to be present in the pig X chromosome. Further investigation of this gene in the pig lineage would aid in determining if multiple copies of the gene are required for encoding these histone proteins.

Studies have suggested H2A-Bbd deficiency in heterochromatic regions of chromosomes; consistent with our findings with the LASTZ hits overlapping genes around the 125Mbp region, often found with lower density of heterochromatin (Chadwick & Willard, 2001; Frönicke & Wienberg, 2001). H2A-Bbd duplications are found to cluster together with high sequence identity having undergone gene conversion leading to differences between mammalian species correlating with our findings – 3 copies at distinct locations and over 99% similarity (Molaro et al., 2018).

The role of histone H2A-Bbd is not entirely understood, particularly in the pig lineage; studies have suggested either a role in spermatogenesis although knockout studies dispute this (Anuar et al., 2019; Molaro et al., 2020). However there was some expression of H2A-Bbd in mice during female meiosis and may play a role in embryonic development (Molaro et al., 2020). Understanding the role of this histone will be important to understand why it is found present with 3 copies in the pig X chromosome along with humans and mice.

Uncharacterised loci are predominantly found to be duplicated in the pig X chromosome however their functions are not yet known

The majority of the duplicated genes are homologous to uncharacterised loci, these genes found at 95Mbp and 45Mbp. The genes at 95Mbp had overlapping hits from 95Mbp and 30Mbp whereas the genes at 45Mbp had overlapping hits from around 47Mbp (Figure 3.27, Figure 3.29, Figure 3.31); correlating with the preliminary LASTZ analysis regions showing inversions and duplications (Figure 3.3). These regions were over 15kb in length suggesting them to be segmental duplications carrying the overlapped genes.

The genes overlapped by these LASTZ alignments were all novel genes found to be long non-coding RNAs - lncRNAs (ENSSSCG00000048704, ENSSSCG00000051484, ENSSSCG00000044950, ENSSSCG00000048218, ENSSSCG00000042077). These genes have no known orthologues on the Ensembl database.

LncRNAs regulate gene expression thereby affecting chromatin modification, transcription and post-transcriptional processing (Mercer et al., 2009). Recruitment of chromatin remodelling complexes by lncRNAs mediates epigenetic changes at specific genomic loci (Mercer et al., 2009), a well-known example previously discussed is *Xist* RNA (see section 1.0 and 4.III). In short, *Xist* RNA recruits proteins involved in chromatin boundary regulation initiating transcriptional silencing through altering gene expression using histone modifications and recruited proteins (Bousard et al., 2019; Zhang & Reinberg, 2001). Recruitment of proteins allows lncRNAs to inhibit or promote gene expression (Statello et al., 2020).

Expression of lncRNAs alters during cell differentiation suggesting they play a role in determining cell fate; particularly as lncRNA expression was linked to gene expression and cell-type-specific gene regulatory functions (Fatica & Bozzoni, 2013). The negative charge alone of lncRNAs can influence chromatin packaging by neutralising the positively charged histone tails affecting their ability to interact with DNA (Dueva et al., 2019). This led to the suggestion lncRNAs function as rapid switches of gene expression (Statello et al., 2020). LncRNAs have also

been seen to regulate expression of maternal/paternal chromosomes using epigenetic markers as a form of imprinting (Fernandes et al., 2019).

The hits overlapping the lncRNA were homologous two different uncharacterised loci in the pig lineage. Firstly, the hits overlapping ENSSSCG00000048704, ENSSSCG00000051484, and ENSSSCG00000044950 (entire coverage and partial coverage) were homologous to the uncharacterised locus LOC110257707. The hits overlapping ENSSSCG00000048218 and ENSSSCG00000042077 (staggered through the gene) were homologous to the uncharacterised locus LOC110257027. The lncRNAs with alignment hits homologous to uncharacterised loci in the pig X chromosome had no phylogenetic branches associated with them. This makes determining their likely function and evolution difficult to determine. The lack of phylogeny and orthologues suggests these loci to be recently evolved in the pig X lineage.

LOC110257707 is found on the pig X chromosome with expression found in in kidney, liver, longissimus dorsi muscle, lung, tissue mixtures, ovary, psoas major muscle, spleen, subcutaneous adipose, and heart. The expression levels are highest in the ovaries however which may suggest this locus to play a role in meiosis or embryogenesis.

LOC110257027 on the other hand is found in chromosome 15 in pigs with expression found in kidney, liver, longissimus dorsi muscle, lung, tissue mixtures, ovary, psoas major muscle, spleen, and subcutaneous adipose. The expression levels are highest in the spleen potentially suggesting it to play a functional role.

Aligning the RNA-seq data for adult pig testis to both loci revealed they are both expressed within the testis. The expression of these loci within gametes could suggest an interesting role of these genes potentially in spermatogenesis & embryogenesis etc. Further investigation to characterise these genes, determine the nature of these loci and their influence on gene expression

would be interesting considering their expression and highly duplicated landscape in the pig X chromosome.

Fragmented repeatmasked LASTZ hits show likely duplication of coagulation factor VIII associated 1 (F8A1)

The gene ENSSSCG00000031635 had fragmented LASTZ hits overlapping its majority in the repeat-masked dataset likely due to the removal of repetitive elements, as the sequence was entirely covered when BLASTing the unmasked FASTA sequence between the LASTZ hit coordinates extracted Ensembl against the gene. ENSSSCG00000031635 is located at ~125Mbp with the overlapping hits found ~8Mbp. According to the Ensembl database ENSSSCG00000031635 is a protein coding gene with 177 orthologues with the majority described to have 1:many orthologues or many:many orthologues (described above). The BLAST results for overlapping hits showed there to be similarity to the predicted coagulation factor VIII associated 1 (F8A1) gene in numerous species.

The F8A1 gene is found in humans, mice, horses, cattle, cats, and pigs as a number of examples. F8A1 is found within intron 22 of the F8 gene, the LASTZ hits BLASTed against the cat and cattle gene showed to primarily overlap the intron 22 protein region (Figure 3.46), further confirming it to be F8A1. In humans the gene is entirely formed of an exon similar to pig annotation (Seefelder et al., 2020). Three copies of this gene are located in the human X chromosome whereas only one fragmented copy was found in our study - showing two total (Seefelder et al., 2020). Copy number of F8A varies between species: mice and cats have only one copy; horses, cattle, and pigs were seen to have two copies. (Seefelder et al., 2020). The human, cattle, and horse copies of this gene share synteny with the pig according to Ensembl, however, the syntenic relationship between the pig, mouse, and cat orthologues was not provided. The

human copies were all found to be around 10kb in size and of very high sequence identity again similar to the findings in the LASTZ alignment (Li et al., 2014).

The evolution of these genes appears to be complex where single exon genes often evolved from intron containing paralogues through DNA-mediated duplication (Zhang et al., 2011). However, Seefelder et al., (2020) suggested this is not the case for F8A1 genes having multi exon orthologues coding for similar proteins therefore it would more likely evolved from gene conversion or double recombination. Roy & Gilbert, (2006) explained double recombination to be where a reverse-transcribed copy of a spliced mRNA transcript deletes adjacent introns; this process is often limited to germline cells to remain inheritable.

Gene duplication can be beneficial where functional proteins can provide a higher quantity of product, for example the F8A1 deficiency is associated with haemophilia A and can be problematic (Zhang, 2003). It is clear the F8A1 gene has evolved in a complex manner between species with varied copy number and locations. There may be a species-specific function of F8A1 leading to the evolution of two copies in the pig X chromosome, although this is unclear at this stage. It would be of interest to determine the function and evolution of the F8A gene family in pigs.

Similarity shown between the LASTZ hit and HSFX variants is likely insignificant

The protein coding gene ENSSSCG00000040153 was overlapped entirely by one LASTZ alignment, through BLAST we found to match HSFX (heat shock transcription factor, X-linked). Ensembl showed ENSSSCG00000040153 to have 44 orthologues with 1:many copy numbers (detailed above) many of which were also HSFX genes, of which the human variants shared synteny. In the BLAST analysis of the overlapping LASTZ hit however, only a few species homologues were returned including the ferret predicted HSFX (Figure 3.39) where the similarity appeared to cover a small portion of the gene.

Heat shock factor (HSF) genes regulate the response to heat stressors and thus regulating gene expression of heat shock genes, known to be conserved within eukaryotes (Wu, 1995). The conservation of the HSFs is seen in the HSF1 orthologue of humans, mice, and cattle sharing ~84% sequence identity (Naidu & Dinkova-Kostova, 2017). HSFs can influence developmental events and cellular processes; such as HSF2, in mice, regulating transcription of multicopy genes on the sex chromosomes e.g. *Ssty* and *Slx* (Barna et al., 2018; Bjö Rk & Sistonen, 2010).

HSFX and HSFY are sex chromosome linked HSF genes seen in species such as the bull, humans, and pigs (Hughes et al., 2020; Skinner et al., 2015). Presence of HSFX and HSFY was suggested on proto-sex chromosomes showing independent evolution and explaining differences in copy number between species (Bhowmick et al., 2007). Apart from the DNA binding domain HSFX and HSFY share little sequence identity further showing their divergence and possibly functional differences (Skinner et al., 2015; Vegesna et al., 2019). HSFY copy number varies greatly between species where their differential amplification highlights independent evolution between species; HSFX on the other hand has reduced amplification observed between species maybe showing conservation of this gene where further studies to determine sequence similarity could validate this (Hughes et al., 2020).

The exact function of HSFX/HSFY is unknown, however, HSFY has testis specific expression in mammals suggesting a role in spermatogenesis (Yue et al., 2014). Human HSFY deletions have been associated with deterioration in spermatogenesis further suggesting it to be functionally significant (Kinoshita et al., 2006; Widlak & Vydra, 2017). In mice, the HSF1 & 2 variants influence spermatogenesis where deletions led to arrest of meiosis and male infertility; possibly where *Slx, Sly,* and *Ssty* require *Hsf* for the correct packaging of chromatin in sperm, without which leads to sperm head abnormalities. (Widlak & Vydra, 2017). There is much less ‘clarity’ on the role of HSFX, except for its high testis expression, again suggesting a potential role in spermatogenesis (Hughes et al., 2020). However, HSFX/Y divergence in sequence

similarity suggests a divergence in functionality, and HSFY is suggested to influence spermatogenesis and male fertility through altering gene expression (Kichine et al., 2012).

In humans the HSFY gene is found to be located within the azoospermic factor b (AZFb) region expressed in Sertoli cells and spermatogenic cells (Kinoshita et al., 2006). Human HSFY is multicopy with at least 2 functional copies, some autosomal pseudogenes, and a duplicated X-linked HSFX (Bhowmick et al., 2006; Kinoshita et al., 2006; Tessari et al., 2004). HSFY is also amplified in the bull with ~79 copies, and ~11 HSFX copies (Hughes et al., 2020). There has also been high amplification of HSFY in pigs with over 100 copies although little evidence of HSFX amplification (Skinner et al., 2015). It would be interesting if there have been duplications of HSFX in pigs as this would follow other species seen with multiple copies of HSFX such as the bull and humans.

At this stage the extent of the similarity of this gene (ENSSSCG00000040153) to HSFX is difficult to determine as Ensembl returns many orthologues not found through the ncbi BLAST alignments – such as the human HSFX1 and HSFX2. The overlaps between the ncbi BLAST results also show only a small proportion of the genes have similarity detected which may not suggest functional significance. This therefore requires further investigation to determine the nature of this similarity before it can be determined if this is a HSFX duplication and then the functional significance within the pig X chromosome can be assessed.

Duplications of genes show evolution and divergence from mammalian conserved genes

Numerous genes and loci have been duplicated in the pig X chromosome, many of which have uncharacterised functions leading to difficulty in determining their functional significance in the pig lineage. Species-specific duplications are a sign of evolution of a species-specific gene

function, adaptation, and divergence of species (Zhang, 2003). Determining whether these duplications are specific to the pig or have specific functions within the pig lineage will aid in understanding the evolution and divergence of the pig genome, and also increase knowledge of the gene content and function in pigs and possibly other mammalian species.

V. No evidence for ampliconic genes on the pig X chromosome shows it to be highly conserved with little or no genomic conflict

Ampliconic genes on the sex chromosomes are often found to be clusters of highly amplified gene families with over 99.9% sequence identity between copies (Skaletsky et al., 2003). This study identified a number of multicopy uncharacterised loci or duplicated genes amongst a vast number of transposable elements, but at lower copy number than would be expected for ampliconic genes such as HSFY and TSPY with between 25-200 copies in mammals such as cattle, humans, and even pigs (Hughes et al., 2020; Oluwole et al., 2017; Skinner et al., 2015); we class the regions of similarity as duplicated genes, of which their functional significance is still unknown.

No ampliconic genes in the pig X chromosome found with testis-specific function contrasting other studies of mammalian amplicons

Mammalian X chromosomes have undergone rapid evolutionary changes leading to acquisition of ampliconic genes, often evolving independently in species, it is said around 10% of human X-linked genes are ampliconic (Vockel et al., 2019). These X-linked amplicons have also been found to have high testis specific expression compared to other tissue types; insinuating a contribution to sperm production however the functions of many of the X amplicons is still unknown (Vockel et al., 2019). The human X-linked ampliconic gene family *Rhox* is seen to have primarily sex-specific expression; the function is not entirely understood, however, it is preferentially expressed in male and female reproductive tracts where mutations in the gene family

have been associated with male infertility and low sperm count (Borgmann et al., 2016; MacLean et al., 2005; Song et al., 2013).

Testis specific expression of X-linked amplicons has also been observed in mice, for example the *Slx* gene family; *Slx* is known to play a role in spermatogenesis (see 1.0) having around 60 copies, with two main variants - *Slx* and *Slxl1 (Sycp3-like X-linked and Slx-like 1)* (Cocquet et al., 2010; Ellis et al., 2011). *Slx* has been directly related to normal mouse sperm development; where deficiency leads to malformation of spermatozoa, defective spermatid elongation, decreased mobility and fertilizing ability (Cocquet et al., 2010).

X-linked amplicons such as those in humans and mice cover a significant proportion of the chromosome and often have testis specific expression and likely play a role in spermatogenesis. There have been no X-linked amplicons with testis specific expression identified in the pig X chromosome, this shows no amplification of testis expressed genes. This may suggest there has been no 'requirement' for amplified X-linked testis gene families in the evolution of the pig genome and this may also suggest there is little conflict between the pig X and Y chromosomes (see below).

Ampliconic gene families in mice exhibit genomic conflict however pig X chromosomes have no amplicons present to play a role in genomic conflict

Genomic conflict as described in 1.0 is well characterised in the mouse sex chromosomes between the ampliconic gene families *Slx* and *Sly;* Often genes in conflict have been newly acquired in a genome, and become co-amplified leading to rapid sequence evolution (Kruger et al., 2019). *Sly* is a recently evolved gene on the mouse Yq acting to suppress sex chromosome transcription in spermatids (Ellis et al., 2011). Having over 100 copies in the MSYq (Y long arm) of mice and it encodes a protein highly expressed in round spermatids; deficiency in *Sly* leads to increased sperm head abnormalities, reducing sperm function and reducing fertility

(Cocquet et al., 2009). The high amplification of spermatid genes act as a counterbalance of the repressive functions of *Sly* to balance the expression levels of the sex-linked genes (Ellis et al., 2011). *Slx* acts opposingly to *Sly* by stimulating XY gene expression in spermatids (Cocquet et al., 2012). This is an example of the genomic conflict between the mouse X and Y chromosomes.

Deletions of either *Slx* or *Sly* results in a sex ratio skew in favour of males or females - respectively within mouse offspring due to their opposing functions (Rathje et al., 2019). For example, *Sly* deficiency leads to an upregulation of X-linked genes including *Slx* (Cocquet et al., 2009). This upregulation of *Slx* leads to a sex ratio skew in favour of females (Bachtrog, 2014; Rathje et al., 2019). Deletions in *Sly* often involve the deletion of a region of the MSYq resulting in sperm head abnormalities and reduced Y bearing sperm motility and in some cases complete removal of *Sly* leads to male infertility (Ward & Burgoyne, 2006). The reverse is also true where *Slx* deletion leads to a skew in favour of male offspring as X bearing sperm have abnormal spermatid elongation, reduced sperm counts and reduced motility (Cocquet et al., 2010; Ellison & Bachtrog, 2019).

This example of conflict results to sex ratio skews and fertility implications of ampliconic genes; these amplicons were co-amplified in the X and Y chromosomes this phenomenon is currently appearing specific to the mouse lineage. Other mammalian species have been seen to have highly amplified Y-linked genes with few or no known X-linked amplicons. Humans have some X-linked amplicons such as *Rhox* or *HSFX,* particularly where *HSFX* has been co-amplified with *HSFY* although there is no evidence for these genes to be in conflict at this time (Larson et al., 2018). This then leads to the question if this conflict is the result of the antagonistic gene amplification or if the conflict led to the amplification.

At this stage, no amplified genes have been identified on the pig X chromosome even with amplified genes present on the Y chromosome such as HSFY (Skinner et al., 2015). This suggests

there are likely no genes in conflict between the chromosomes and the Y chromosome amplifications may be acting to counteract the degradation of the Y chromosome during sex chromosome evolution (see section 1.0). Further investigation into the relationship between genomic conflict and amplification of genes may be of interest to determine the cause of the highly amplified mouse genes and understanding genomic conflict in greater detail. It may also be of interest to determine whether the lack of amplicons on the pig X chromosome is unique to pigs or likely a more general feature of mammalian X chromosomes.

Ampliconic genes appear to be more common on lesser conserved sex chromosomes such as that of mice

As discussed in section 1.II and 1.0, the reduced recombination and independent evolution of the X and Y chromosomes gave rise to an increase in sequence amplifications (Lucotte et al., 2018; Skinner et al., 2016). This was triggered by inversions on the Y chromosome likely in order to counterbalance gene loss and degradation of the MSY (Hughes et al., 2012; Soh et al., 2014). Y linked amplicons are a common feature of mammalian sex chromosomes and are independently acquired prior to amplification regardless of their similar patterns of expression and influence on spermatogenesis (Larson et al., 2018).

Mammalian X-linked amplicons however are less common as seen where no evidence for ampliconic genes was found on the pig X chromosome. There has been much less re-arrangement in X chromosomes compared to Y chromosomes, and even some autosomes (Ohno, 1966). Comparisons of human and mouse X linked genes showed single-copy genes were conserved from a common ancestor with over 95% of copies shared between species; however exceptions of this conservation were seen when comparing ampliconic genes where orthologues were only seen with 31% of human copies and with 22% of mouse copies (Mueller et al., 2013). Many X ampliconic

genes are also documented to be independently acquired since the divergence from a common ancestor, between humans and mice, with testis specific expression (Mueller et al., 2013).

Further comparisons between mouse, human, and horse X chromosomes revealed large syntenic relationships between human and horse gene orders, however little similarity shared between these species and mice (Raudsepp et al., 2002; Sandstedt & Tucker, 2004). This highlights the high levels of re-arrangement in the mouse X chromosome possibly explaining the much greater ampliconic content compared to other mammals. The mouse amplicons may have also been exasperated in copy number due to the evolutionary arms race between antagonistic genes in the mouse sex chromosomes (Larson et al., 2018). The human X has some co-amplified genes present however nowhere near the extent of mice; this further confirms amplicons are less common in highly conserved chromosomes. Comparisons between the human and pig X chromosomes showed few rearrangements had occurred (Mccoard et al., 2002).

This study found no evidence of amplicon enrichment on the pig X chromosome; the pig X is highly conserved supporting the suggestion amplicons are more common on less conserved chromosomes. This may be due to rearrangements causing gene amplifications however this is unknown as amplifications may be the cause of structural rearrangements. It would be of interest to determine the relationship between gene amplification and chromosomal rearrangements; thus, informing why conserved X chromosomes have fewer ampliconic genes, and to understand the high occurrence of amplicons in the mouse X chromosome.

VI. Conclusions and future perspectives

This study identified unannotated repetitive elements, gene, and segmental duplications, yet there has been no evidence for enrichment of ampliconic genes on the pig X chromosome. Embedded L1 sequences suggest a recent transposition occurred in the pig lineage where the sequence appears to be pig specific. This may have altered gene expression and gene function in

the pig lineage, but the active state and functional significance of these L1s still needs to be determined. Evidence of segmental and gene duplications in the pig X may be a sign of independent divergence in the pig lineage from the common ancestor for mammalian species; in order to acquire a species-specific function. Many of the characterised genes had orthologues in other species both with multiple copies of the gene; however, the uncharacterised loci and lncRNAs showed less similarity. While these lncRNAs are expressed in multiple tissues, more research is required to determine their impact on gene expression. These insertions and duplications suggest recent evolution in the pig genome to acquire species specific genes and functions. To validate this, there should be further investigation into the expression of these genes and genes within these regions. It would also be beneficial to have a physical validation of the sequence copy number through methods such as quantitative PCR.

At this stage, we can improve the current annotation of the pig X chromosome with the L1 sequences which were previously unannotated. There have also been genes and loci of interest identified for further investigation to determine their evolution and function in the pig lineage. This analysis was performed on the reference genome, which is primarily formed of Duroc pig data, therefore it may be worthwhile investigating other pig breeds to increase the scope of the investigation. Finally, we have not seen evidence of ampliconic genes in the pig X chromosome, raising questions to be investigated regarding the nature of ampliconic genes and why they have occurred in less conserved sex chromosomes such as in the mouse.

5. References

Genome - Assembly - NCBI. (n.d.). Retrieved 17 June 2021, from https://www.ncbi.nlm.nih.gov/assembly

Abbott, J. K., Nordén, A. K., & Hansson, B. (2017). Sex chromosome evolution: historical insights and future perspectives. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1854), 20162806. https://doi.org/10.1098/rspb.2016.2806

Acquaviva, L., Boekhout, M., Karasu, M., Brick, K., Pratto, F., Li, T., Overbeek, M., Kauppi, L., Camrini-Otero, D., Jasin, M., & Keeney, S. (2020). Ensuring meiotic DNA break formation in the mouse pseudoautosomal region. *Nature*, *582*(7812), 426–431.

Adega, F., Chaves, R., & Guedes-Pinto, H. (2005). Chromosome Restriction Enzyme Digestion in Domestic Pig (Sus scrofa) Constitutive heterochromatin arrangement. *Genes & Genetic Systems*, *80*(1), 49–56.

Ai, H., Chen, H., & Mao, L. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, *47*(3), 217–225. https://doi.org/10.1038/ng.3199

Alamro, H., Alzamel, M., Iliopoulos, C. S., Pissis, S. P., & Watts, S. (2021). IUPACpal: efficient identification of inverted repeats in IUPAC-encoded DNA sequences. *BMC Bioinformatics*, *22*(1), 1–12. https://doi.org/10.1186/S12859-021-03983-2

Anuar, N., Kurscheid, S., Field, M., Zhang, L., Rebar, E., Gregory, P., Buchou, T., Bowles, J., Koopman, P., Tremethick, D., & Soboleva, T. (2019). Gene editing of the multi-copy H2A.B gene and its importance for fertility. *Genome Biology*, *20*(1). https://doi.org/10.1186/S13059-019-1633-3

Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. In *Annual Review of Animal Biosciences* (Vol. 7, pp. 41–64). Annual Reviews Inc. https://doi.org/10.1146/annurev-animal-020518-115005

Arslan, A. N., Harvey Greenberg, by J., & Editor, G. (2004). Dynamic Programming Based Approximation Algorithms for Sequence Alignment with Constraints. *INFORMS Journal on Computing*, *16*(4), 441–458. https://doi.org/10.1287/ijoc.1040.0097

Arvid Agren, J. (2016). Selfish genetic elements and the gene's-eye view of evolution. *Current Zoology*, *62*(6), 659–665. https://doi.org/10.1093/cz/zow102

Bachtrog, D. (2006). A dynamic view of sex chromosome evolution. *Current Opinion in Genetics and Development*, *16*(6), 578–585. https://doi.org/10.1016/j.gde.2006.10.007

Bachtrog, D. (2013). Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. In *Nature Reviews Genetics* (Vol. 14, Issue 2, pp. 113–124). NIH Public Access. https://doi.org/10.1038/nrg3366

Bachtrog, D. (2014). Signs of Genomic Battles in Mouse Sex Chromosomes. *Cell*, *159*(4), 716–718. https://doi.org/https://doi.org/10.1016/j.cell.2014.10.036.

Bachtrog, D., Kirkpatrick, M., Mank, J. E., McDaniel, S. F., Pires, J. C., Rice, W. R., & Valenzuela, N. (2011). Are all sex chromosomes created equal? In *Trends in Genetics* (Vol. 27, Issue 9, pp. 350–357). Elsevier Ltd. https://doi.org/10.1016/j.tig.2011.05.005

Bakalov, V. K., Cheng, C., Zhou, J., & Bondy, C. A. (2009). X-Chromosome Gene Dosage and the Risk of Diabetes in Turner Syndrome. *The Journal of Clinical Endocrinology & Metabolism*, *94*(9), 3289–3296. https://doi.org/10.1210/jc.2009-0384

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* , *6*(1), 1–6. https://doi.org/10.1186/S13100-015-0041-9

Barna, J., Csermely, P., & Tibor Vellai, ·. (2018). Roles of heat shock factor 1 beyond the heat shock response. *Cellular and Molecular Life Sciences*, *75*, 2897–2916. https://doi.org/10.1007/s00018-018-2836-6

Baroiller, J. F., & D'Cotta, H. (2016). The Reversible Sex of Gonochoristic Fish: Insights and Consequences. In *Sexual Development* (Vol. 10, Issues 5–6, pp. 242–266). S. Karger AG. https://doi.org/10.1159/000452362

Batzoglou, S. (2005). The many faces of sequence alignment. *Briefings in Bioinformatics*, *6*(1), 6–22. https://doi.org/10.1093/bib/6.1.6

Baum, M. J. (2012). Contribution of pheromones processed by the main olfactory system to mate recognition in female mammals. *Frontiers in Neuroanatomy*, *6*, 20. https://doi.org/10.3389/fnana.2012.00020

Bell, G. (1982). *The masterpiece of nature: the evolution and genetics of sexuality*. Cambridge university press.

Bellott, D., Hughes, J., Skaletsky, H., Brown, L., Pyntikova, T., Cho, T. J., Koutseva, N., Zaghlul, S., Graves, T., Rock, S., Kremitzki, C., Fulton, R. S., Dugan, S., Ding, Y., Morton, D., Khan, Z., Lewis, L., Buhay, C., Wang, Q., … Page, D. C. (2014). Mammalian y chromosomes retain widely expressed dosage-sensitive regulators. *Nature*, *508*(7497), 494–499. https://doi.org/10.1038/nature13206

Bellott, D. W., & Page, D. C. (2009). Reconstructing the Evolution of Vertebrate Sex Chromosomes. *Cold Spring Harbor Symposia on Quantitative Biology*, *74*, 345-353. https://doi.org/10.1101/sqb.2009.74.048

Bendixen, E., Danielsen, M., Larsen, K., & Bendixen, C. (2010). Advances in porcine genomics and proteomicsc-a toolbox for developing the pig as a model organism for molecular

biomedical research. *Briefings in Functional Genomics and Proteomics*, *9*(3), 208–219. https://doi.org/10.1093/bfgp/elq004

Berletch, J. B., Ma, W., Yang, F., Shendure, J., Noble, W. S., Disteche, C. M., & Deng, X. (2015). Escape from X Inactivation Varies in Mouse Tissues. *PLoS Genetics*, *11*(3), 1005079. https://doi.org/10.1371/journal.pgen.1005079

Berletch, J. B., Yang, F., & Disteche, C. M. (2010). Escape from X inactivation in mice and humans. In *Genome Biology* (Vol. 11, Issue 6, pp. 1–7). BioMed Central. https://doi.org/10.1186/gb-2010-11-6-213

Berletch, J. B., Yang, F., Xu, J., Carrel, L., & Disteche, C. M. (2011). Genes that escape from X inactivation. *Human Genetics*, *130*(2), 237–245. https://doi.org/10.1007/s00439-011-1011-z

Berta, P., Hawkins, B., Sinclair, A. H., Taylor, A., Griffiths, B., Goodfellow, P., & Fellous, M. (1990). Genetic evidence equating SRY and the testis-determining factor. *Nature*, *348*(6300), 448–450.

Beukeboom, L., & Perrin, N. (2014). *The Evolution of Sex Determination*. Oxford University Press.

Bhowmick, B. K., Satta, Y., & Takahata, N. (2007). The origin and evolution of human ampliconic gene families and ampliconic structure. *Genome Research*, *17*(4), 441–450. https://doi.org/10.1101/GR.5734907

Bhowmick, B., Takahata, N., Watanabe, M., & Satta, Y. (2006). Comparative analysis of human masculinity. *Genetics and Molecular Research*, *5*(4), 696–712.

Bjö Rk, J. K., & Sistonen, L. (2010). Regulation of the members of the mammalian heat shock factor family. *The FEBS Journal* , *277*(20), 4126–4139. https://doi.org/10.1111/j.1742-4658.2010.07828.x

Boissinot, S., Entezam, A., & Furano, A. V. (2001). Selection Against Deleterious LINE-1-Containing Loci in the Human Lineage. *Molecular Biology and Evolution*, *18*(6), 926–935. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A003893

Bonneau, M., & Weiler, U. (2019). Pros and cons of alternatives to piglet castration: Welfare, boar taint, and other meat quality traits. In *Animals* (Vol. 9, Issue 11, p. 884). MDPI AG. https://doi.org/10.3390/ani9110884

Borgmann, J., Tüttelmann, F., Dworniczak, B., Röpke, A., Song, H.-W., Kliesch, S., Wilkinson, M. F., Laurentino, S., & Gromoll, J. (2016). The human RHOX gene cluster: target genes and functional analysis of gene variants in infertile men. *Human Molecular Genetics*, *25*(22), 4898. https://doi.org/10.1093/HMG/DDW313

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1577-z

Bousard, A., Raposo, A. C., Żylicz, J. J., Picard, C., Pires, V. B., Qi, Y., Gil, C., Syx, L., Chang, H. Y., Heard, E., & da Rocha, S. T. (2019). The role of Xist -mediated Polycomb recruitment in the initiation of X-chromosome inactivation. *EMBO Reports*, *20*(10), e48019. https://doi.org/10.15252/embr.201948019

Brown, C., Ballabio, A., Rupert, J., Lafreniere, R., Grompe, M., Tonlorenzi, R., & Willard, H. F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, *349*(6304), 38–44.

Brudno, M., Chapman, M., Göttgens, B., Batzoglou, S., & Morgenstern, B. (2003). Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, *4*(1), 1–11. https://doi.org/10.1186/1471-2105-4-66

Burt, A., & Trivers, R. (2009). *Genes in Conflict: The Biology of Selfish Genetic Elements*. Harvard university press.

Butcher, D. (1995). Muller's Ratchet, Epistasis and Mutation Effects. *Genetics*, *141*(1), 431–437.

Canzar, S., & Salzberg, S. L. (2015). Short Read Mapping: An Algorithmic Tour. *Proceedings of the IEEE*, *105*(3), 436–458. https://doi.org/10.1109/JPROC.2015.2455551

Cao, L. A. (2012). *Elucidating the Roles of XPMC2H and Spin3 in Mitosis*.

Carrel, L., & Brown, C. J. (2017). When the Lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1733), 20160355. https://doi.org/10.1098/rstb.2016.0355

Carroll, L. S., Meagher, S., Morrison, L., Penn, D. J., & Potts, W. K. (2004). Fitness effects of a selfish gene (the Mus t complex) are revealed in an ecological context. *Evolution*, *58*(6), 1318–1328. https://doi.org/10.1111/j.0014-3820.2004.tb01710.x

Carvalho, A. B., Sampaio, M. C., Varandas, F. R., & Klaczko, L. B. (1998). An Experimental Demonstration of Fisher's Principle: Evolution of Sexual Proportion by Natural Selection. *Genetics*, *148*(2), 719–731.

Chadwick, B. P., & Willard, H. F. (2001). A Novel Chromatin Protein, Distantly Related to Histone H2a, Is Largely Excluded from the Inactive X Chromosome. *Journal of Cell Biology*, *152*(2), 375–384. https://doi.org/10.1083/JCB.152.2.375

Chain, P., Kurtz, S., Ohlebusch, E., & Slezak, T. (2003). An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. *Briefings in Bioinformatics*, *4*(2), 105–123. https://doi.org/10.1093/bib/4.2.105

Chaligné, R., & Heard, E. (2014). X-chromosome inactivation in development and cancer. In *FEBS Letters* (Vol. 588, Issue 15, pp. 2514–2522). Elsevier.

https://doi.org/10.1016/j.febslet.2014.06.023

Charlesworth, B. (1978). Model for evolution of Y chromosomes and dosage compensation. *Proceedings of the National Academy of Sciences*, *75*(11), 5618–5622. https://doi.org/10.1073/pnas.75.11.5618

Charlesworth, Brian. (1991). The evolution of sex chromosomes. *Science*, *251*(4997), 1030–1033. https://doi.org/10.1126/science.1998119

Charlesworth, Brian. (1996). The evolution of chromosomal sex determination and dosage compensation. *Current Biology*, *6*(2), 149–162. https://doi.org/10.1016/S0960-9822(02)00448-7

Charlesworth, Brian. (2002). *The evolution of chromosomal sex determination*.

Charlesworth, Brian, & Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London*, *355*(1403), 1563–1572. https://doi.org/10.1098/rstb.2000.0717

Charlesworth, D. (2017). Evolution of recombination rates between sex chromosomes. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 372, Issue 1736, p. 20160456). Royal Society Publishing. https://doi.org/10.1098/rstb.2016.0456

Chen, G., Schell, J. P., Aguila Benitez, J., Petropoulos, S., Yilmaz, M., Reinius, B., Alekseenko, Z., Shi, L., Hedlund, E., Lanner, F., Sandberg, R., & Deng, Q. (2016). Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Research*, *26*(10), 1342–1354. https://doi.org/10.1101/gr.201954.115

Chen, K., Baxter, T., Muir, W. M., Groenen, M. A., & Schook, L. B. (2007). Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). In *International Journal of Biological Sciences* (Vol. 3, Issue 3, pp. 153–165). Ivyspring International Publisher.

https://doi.org/10.7150/ijbs.3.153

Chew, G.-L., Bleakley, M., Bradley, R. K., Malik, H. S., Henikoff, S., Molaro, A., & Sarthy, J. (2021). Short H2A histone variants are expressed in cancer. *Nature Communications*, *12*(1), 1–9. https://doi.org/10.1038/s41467-020-20707-x

Chow, J. C., Ciaudo, C., Fazzari, M. J., Mise, N., Servant, N., Glass, J. L., Attreed, M., Avner, P., Wutz, A., Barillot, E., Greally, J. M., Voinnet, O., & Heard, E. (2010). LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell*, *141*(6), 956–969. https://doi.org/10.1016/j.cell.2010.04.042

Clemson, C. M., McNeil, J. A., Willard, H. F., & Lawrence, J. B. (1996). XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear/chromosome structure. *Journal of Cell Biology*, *132*(3), 259–275. https://doi.org/10.1083/jcb.132.3.259

Cocquet, J., Ellis, P. J. I., Mahadevaiah, S. K., Affara, N. A., Vaiman, D., & Burgoyne, P. S. (2012). A Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse. *PLoS Genetics*, *8*(9), e1002900. https://doi.org/10.1371/journal.pgen.1002900

Cocquet, J., Ellis, P. J. I., Yamauchi, Y., Riel, J. M., Karacs, T. P. S., Ine Rattigan, A. ′, Ojarikre, O. A., Affara, N. A., Ward, M. A., & Burgoyne, P. S. (2010). Deficiency in the Multicopy Sycp3-Like X-Linked Genes Slx and Slxl1 Causes Major Defects in Spermatid Differentiation. *Molecular Biology of the Cell*, *21*(20), 3497–3505. https://doi.org/10.1091/mbc.E10

Cocquet, J., Ellis, P., Yamauchi, Y., Mahadevaiah, S., Affara, N. A., & Ward, M. (2009). The Multicopy Gene Sly Represses the Sex Chromosomes in the Male Mouse Germline after Meiosis. *PLoS Biology*, *7*(11), p.e1000244. https://doi.org/https://doi.org/10.1371/journal.pbio.1000244

Colaco, S., & Modi, D. (2018). Genetics of the human Y chromosome and its association with male infertility. In *Reproductive Biology and Endocrinology* (Vol. 16, Issue 1, pp. 1–24). BioMed Central Ltd. https://doi.org/10.1186/s12958-018-0330-5

Comeron, J. M. (2005). Intragenic Hill-Robertson Interference Influences Selection Intensity on Synonymous Mutations in Drosophila. *Molecular Biology and Evolution*, *22*(12), 2519–2530. https://doi.org/10.1093/molbev/msi246

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* , *10*(10), 691–703. https://doi.org/10.1038/nrg2640

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F., & Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature*, *508*(7497), 488–493. https://doi.org/10.1038/nature13151

Crespi, B., & Nosil, P. (2013). Conflictual speciation: species formation via genomic conflict. *Trends in Ecology & Evolution*, *28*(1), 48–57. https://doi.org/10.1016/j.tree.2012.08.015

da Silva, J., & Galbraith, J. D. (2017). Hill-Robertson interference maintained by Red Queen dynamics favours the evolution of sex. *Journal of Evolutionary Biology*, *30*(5), 994–1010. https://doi.org/10.1111/jeb.13068

Dang, V. T., Kassahn, K. S., Marcos, A. E., & Ragan, M. A. (2008). Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *European Journal of Human Genetics*, *16*(11), 1350–1357. https://doi.org/10.1038/ejhg.2008.111

Darmon, E., Eykelenboom, J. K., Lincker, F., Jones, L. H., White, M., Okely, E., Blackwood, J. K., & Leach, D. R. (2010). E. coli SbcCD and RecA Control Chromosomal Rearrangement Induced by an Interrupted Palindrome. *Molecular Cell*, *39*(1), 59–70. https://doi.org/10.1016/J.MOLCEL.2010.06.011

Dechaud, C., Volff, J. N., Schartl, M., & Naville, M. (2019). Sex and the TEs: Transposable elements in sexual development and function in animals. In *Mobile DNA* (Vol. 10, Issue 1, pp. 1–15). BioMed Central Ltd. https://doi.org/10.1186/s13100-019-0185-0

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, *27*(11), 2369–2376. https://doi.org/10.1093/nar/27.11.2369

Deng, X., Berletch, J. B., Ma, W., Nguyen, D. K., Hiatt, J. B., Noble, W. S., Shendure, J., & Disteche, C. M. (2013). Mammalian X upregulation is associated with enhanced transcription initiation, RNA half-life, and MOF-mediated H4K16 acetylation. *Developmental Cell*, *25*(1), 55–68. https://doi.org/10.1016/j.devcel.2013.01.028

Deng, X., Nguyen, D. K., Hansen, R. S., Van Dyke, D. L., Gartler, S. M., & Disteche, C. M. (2009). Dosage Regulation of the Active X Chromosome in Human Triploid Cells. *PLoS Genetics*, *5*(12), e1000751. https://doi.org/10.1371/journal.pgen.1000751

Devlin, R. H., & Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture*, *208*(3–4), 191–364.

Dewey, C. N. (2019). Whole-genome alignment. In *Methods in Molecular Biology* (Vol. 1910, pp. 121–147). Humana Press Inc. https://doi.org/10.1007/978-1-4939-9074-0_4

Disteche, C. M. (1999). Escapees on the X chromosome. In *Proceedings of the National Academy of Sciences* (Vol. 96, Issue 25, pp. 14180–14182). National Academy of Sciences. https://doi.org/10.1073/pnas.96.25.14180

Disteche, C. M. (2016). Dosage compensation of the sex chromosomes and autosomes. In *Seminars in Cell and Developmental Biology* (Vol. 56, pp. 9–18). Academic Press.

https://doi.org/10.1016/j.semcdb.2016.04.013

Disteche, C. M., & Berletch, J. B. (2015). X-chromosome inactivation and escape. *Journal of Genetics*, *94*(4), 591–599.

Dueva, R., Akopyan, K., Pederiva, C., Trevisan, D., Dhanjal, S., Lindqvist, A., & Farnebo, M. (2019). Neutralization of the Positive Charges on Histone Tails by RNA Promotes an Open Chromatin Structure. *Cell Chemical Biology*, *26*(10), 1436-1449.e5. https://doi.org/10.1016/J.CHEMBIOL.2019.08.002

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Dutheil, J. Y., Munch, K., Nam, K., Mailund, T., & Schierup, M. H. (2015). Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. *PLOS Genetics*, *11*(8), e1005451. https://doi.org/10.1371/journal.pgen.1005451

Eddy, S. (2004). What is dynamic programming? *Nature Biotechnology*, *22*(7), 909–910.

Ellis, P. J. I., Bacon, J., & Affara, N. A. (2011). Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids? *Human Molecular Genetics*, *20*(15), pp.3010-3021. https://doi.org/10.1093/hmg/ddr204

Ellison, C., & Bachtrog, D. (2019). Recurrent gene co-amplification on Drosophila X and Y chromosomes. *PLoS Genetics*, *15*(7), e1008251. https://doi.org/10.1371/journal.pgen.1008251

Emms, D., & Kelly, S. (2021). *SHOOT.bio*. https://shoot.bio/

Ensembl genome browser 101. (n.d.). Retrieved 21 November 2020, from http://www.ensembl.org/index.html

Ezaz, T., Stiglec, R., Veyrunes, F., & Marshall Graves, J. A. (2006). Relationships between Vertebrate ZW and XY Sex Chromosome Systems. In *Current Biology* (Vol. 16, Issue 17, pp. R736–R743). Cell Press. https://doi.org/10.1016/j.cub.2006.08.021

Fang, H., Disteche, C. M., & Berletch, J. B. (2019). X Inactivation and Escape: Epigenetic and Structural Features. In *Frontiers in Cell and Developmental Biology* (Vol. 7, p. 219). Frontiers Media S.A. https://doi.org/10.3389/fcell.2019.00219

Fatica, A., & Bozzoni, I. (2013). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics* , *15*(1), 7–21. https://doi.org/10.1038/nrg3606

Fernandes, J. C. R., Acuña, S. M., Aoki, J. I., Floeter-Winter, L. M., & Muxel, S. M. (2019). Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. *Non-Coding RNA* , *5*(1), 17. https://doi.org/10.3390/NCRNA5010017

Fisher, R. A. (1935). The sheltering of lethals. *American Naturalist*, *69*(724), 446–455.

Fisher, R. A. (1958). *Retrospect of the Criticisms of the Theory of Natural Selection*.

Foster, J. W., Graves, J. A. M., & Cooper, D. W. (1994). An SRY-related sequence on the marsupial X chromosome: Implications for the evolution of the mammalian testis-determining gene. In *Proceedings of the national academy of sciences* (Vol. 91).

Frank, S. A., & Crespi, B. J. (2011). Pathology from evolutionary conflict, with a theory of X chromosome versus autosome conflict over sexually antagonistic traits. *Proceedings of the National Academy of Sciences*, *108*(2), 10886–10893. https://doi.org/10.1073/pnas.1100921108

Fridolfsson, A. K., Cheng, H., Copeland, N. G., Jenkins, N. A., Liu, H. C., Raudsepp, T., Woodage, T., Chowdhary, B., Halverson, J., & Ellegren, H. (1998). Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proceedings of the National Academy of*

Sciences, *95*(14), 8147–8152. https://doi.org/10.1073/pnas.95.14.8147

Frönicke, L., & Wienberg, J. (2001). Comparative chromosome painting defines the high rate of karyotype changes between pigs and bovids. *Mammalian Genome 2001 12:6*, *12*(6), 442–449. https://doi.org/10.1007/S003350010288

Gabriel, W., Lynch, M., & Bürger, R. (1993). Muller's ratchet and mutational meltdowns. *Evolution*, *47*(6), 1744–1757. https://doi.org/10.1111/j.1558-5646.1993.tb01266.x

Ganapathiraju, M. K., Subramanian, S., Chaparala, S., & Karunakaran, K. B. (2020). A reference catalog of DNA palindromes in the human genome and their variations in 1000 Genomes. *Human Genome Variation* , *7*(1), 1–12. https://doi.org/10.1038/s41439-020-00127-5

Gao, K., & Miller, J. (2014). Human-chimpanzee alignment: Ortholog exponentials and paralog power laws. *Computational Biology and Chemistry*, *53*(PA), 59–70. https://doi.org/10.1016/j.compbiolchem.2014.08.010

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. In *Development (Cambridge)* (Vol. 143, Issue 22, pp. 4101–4114). Company of Biologists Ltd. https://doi.org/10.1242/dev.132639

Gardner, A., & Welch, J. J. (2011). A formal theory of the selfish gene. *Journal of Evolutionary Biology*, *24*(8), 1801–1813. https://doi.org/10.1111/j.1420-9101.2011.02310.x

Gardner, Andy, & Francisco, U. (2017). The meaning of intragenomic conflict. *Nature Ecology and Evolution*, *1*(12), 1807–1815.

geom_segment function - RDocumentation. (n.d.). R Documentation. Retrieved 1 July 2021, from https://www.rdocumentation.org/packages/ggplot2/versions/3.3.4/topics/geom_segment

Gibson, C. E., Boodhansingh, K. E., Li, C., Conlin, L., Chen, P., Becker, S. A., Bhatti, T., Bamba, V., Adzick, N. S., De Leon, D. D., Ganguly, A., & Stanley, C. A. (2018). Congenital

Hyperinsulinism in Infants with Turner Syndrome: Possible Association with Monosomy X and KDM6A Haploinsufficiency. *Hormone Research in Paediatrics*, *89*(6), 413–422. https://doi.org/10.1159/000488347

Gil-Fernández, A., Saunders, P. A., Martín-Ruiz, M., Ribagorda, M., López-Jiménez, P., Jeffries, D. L., Parra, M. T., Viera, A., Rufas, J. S., Perrin, N., Veyrunes, F., & Page, J. (2020). Meiosis reveals the early steps in the evolution of a neo-XY sex chromosome pair in the African pygmy mouse Mus minutoides. *PLOS Genetics*, *16*(11), e1008959. https://doi.org/10.1371/journal.pgen.1008959

Gordo, I., & Charlesworth, B. (2000). The Degeneration of Asexual Haploid Populations and the Speed of Muller's Ratchet. *Genetics*, *154*(3), 1379–1387.

Gordo, I., Navarro, A., & Charlesworth, B. (2002). Muller's Ratchet and the Pattern of Variation at a Neutral Locus. *Genetics*, *161*(2), 835–848.

Graves, J. A. M. (2002). The rise and fall of SRY. In *Trends in Genetics* (Vol. 18, Issue 5, pp. 259–264). Elsevier Current Trends. https://doi.org/10.1016/S0168-9525(02)02666-5

Graves, J. M. (2006). Sex chromosome specialization and degeneration in mammals. In *Cell* (Vol. 124, Issue 5, pp. 901–914). Elsevier. https://doi.org/10.1016/j.cell.2006.02.024

Groenen, M. A. M. (2016). A decade of pig genome sequencing: A window on pig domestication and evolution. In *Genetics Selection Evolution* (Vol. 48, Issue 1, pp. 1–9). BioMed Central Ltd. https://doi.org/10.1186/s12711-016-0204-2

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H. J., Li, S., Larkin, D. M., Kim, H., Frantz, L. A. F., Caccamo, M., Ahn, H., Aken, B. L., Anselmo, A., Anthon, C., … Schook, L. B. (2012). Analyses of pig genomes provide insight into porcine demography and evolution.

Nature, *491*(7424), 393–398. https://doi.org/10.1038/nature11622

Gullett, E. A., Partlow, G. D., Fisher, K. R. S., Halina, W. G., & Squires, E. J. (1993). Effect of pig sex on consumer ratings for pork chops and bacon. *Food Quality and Preference*, *4*(4), 201–205. https://doi.org/10.1016/0950-3293(93)90163-Z

Gummere, G. R., Mccormick' And, P. J., & Bennett, D. (1986). The influence of genetic background and the transmission ratio distortion in mice homologous chromosome 17 on t-haplotype. *Genetics*, *114*(1), 235–245.

Gupta, V., Parisi, M., Sturgill, D., Nuttall, R., Doctolero, M., Dudko, O. K., Malley, J. D., Eastman, P. S., & Oliver, B. (2006). Global analysis of X-chromosome dosage compensation. *Journal of Biology*, *5*(1), 1–22. https://doi.org/10.1186/jbiol30

Gustavsson, L. (1988). Standard karyotype of the domestic pig Committee: for the Standardized Karyotype of the Domestic Pig. *Hereditas*, *109*(2), 151–157.

Haig, D. (2006). Intragenomic politics. *Cytogenetic and Genome Research*, *113*(1–4), 68–74. https://doi.org/10.1159/000090816

Hamilton, W. (1967). Extraordinary sex ratios. *Science*, *156*(3774), 477–488.

Hamilton, W. D. (1964). The Genetical Evolution of Social Behaviour. II. In *Journal of theoretical biology* (Vol. 7).

Hansen-Melander, E., & Melander, Y. (1974). The karyotype of the pig. *Hereditas*, *77*(1), 149–158. https://doi.org/10.1111/J.1601-5223.1974.TB01358.X

Harris, B. (2010). *LASTZ*. http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html#stages_detail

Harris, R. S. (2007). *Improved Pairwise Alignment of Genomic DNA*.

Heard, E., Chaumeil, J., Masui, O., & Okamoto, I. (2004). Mammalian X-Chromosome Inactivation: An Epigenetics Paradigm. In *Wutz and Cold Spring Harbor Symposia on Quantitative Biology: Vol. LXIX*. Cold Spring Harbor Laboratory Press.

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, *29*(10), 1341–1342. https://doi.org/10.1093/bioinformatics/btt128

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *3*(8), 269–294. https://doi.org/10.1017/S001667230800949X

Hitchcock, T. J., & Gardner, A. (2020). A gene's-eye view of sexual antagonism. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1932), 20201633. https://doi.org/10.1098/rspb.2020.1633

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., & Xu, A. (2012). HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Research*, *22*(8), 1581–1588. https://doi.org/10.1101/gr.133652.111

Huang, Y., & Zhang, L. (2004). Rapid and sensitive dot-matrix methods for genome analysis. *Bioinformatics*, *20*(4), 460–466. https://doi.org/10.1093/bioinformatics/btg429

Hughes, J. F., & Page, D. C. (2015). The Biology and Evolution of Mammalian Y Chromosomes. *Annual Review of Genetics*, *49*(1), 507–527. https://doi.org/10.1146/annurev-genet-112414-055311

Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Graves, T., Fulton, R. S., Dugan, S., Ding, Y., Buhay, C. J., Kremitzki, C., Wang, Q., Shen, H., Holder, M., Villasana, D., Nazareth, L. V., Cree, A., Courtney, L., Veizer, J., Kotkiewicz, H., … Page, D. C. (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus y

chromosomes. *Nature*, *483*(7387), 82–87. https://doi.org/10.1038/nature10843

Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., Van Daalen, S. K. M., Minx, P. J., Fulton, R. S., McGrath, S. D., Locke, D. P., Friedman, C., Trask, B. J., Mardis, E. R., Warren, W. C., Repping, S., Rozen, S., Wilson, R. K., & Page, D. C. (2010). Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. *Nature*, *463*(7280), 536–539. https://doi.org/10.1038/nature08700

Hughes, J. F., Skaletsky, H., Pyntikova, T., Koutseva, N., Raudsepp, T., Brown, L. G., Bellott, D. W., Cho, T.-J., Dugan-Rocha, S., Khan, Z., Kremitzki, C., Fronick, C., Graves-Lindsay, T. A., Fulton, L., Warren, W. C., Wilson, R. K., Owens, E., Womack, J. E., Murphy, W. J., … Page, D. C. (2020). Sequence analysis in Bos taurus reveals pervasiveness of X–Y arms races in mammalian lineages . *Genome Research*, *30*(12), 1716–1726. https://doi.org/10.1101/gr.269902.120

Hughes, J. F., Skaletsky, H., Pyntikova, T., Minx, P. J., Graves, T., Rozen, S., Wilson, R. K., & Page, D. C. (2005). Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature*, *437*(7055), 100–103. https://doi.org/10.1038/nature04101

Iacolina, L., Brajković, V., Canu, A., Šprem, N., Cubric-Curik, V., Fontanesi, L., Saarma, U., Apollonio, M., & Scandura, M. (2016). Novel Y-chromosome short tandem repeats in *Sus scrofa* and their variation in European wild boar and domestic pig populations. *Animal Genetics*, *47*(6), 682–690. https://doi.org/10.1111/age.12483

Irwin, D. E. (2018). Sex chromosomes and speciation in birds and other ZW systems. *Molecular Ecology*, *27*(19), 3831–3851. https://doi.org/10.1111/mec.14537

Ishibashi, T., Li, A., Eirín-López, J. M., Zhao, M., Missiaen, K., Abbott, D. W., Meistrich, M., Hendzel, M. J., & Ausió, J. (2010). H2A.Bbd: an X-chromosome-encoded histone involved

in mammalian spermiogenesis. *Nucleic Acids Research*, *38*(6), 1780–1789. https://doi.org/10.1093/NAR/GKP1129

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., & Adelson, D. L. (2016). LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. *Genome Biology and Evolution*, *8*(11), 3301–3322. https://doi.org/10.1093/GBE/EVW243

Jackson, E. K., Bellott, D. W., Cho, T.-J., Skaletsky, H., Hughes, J. F., Pyntikova, T., & Page, D. C. (2020). Large palindromes on the primate X Chromosome are preserved by natural selection. *Genome Research*, *31*(8), 1337–1352. https://doi.org/10.1101/GR.275188.120

Janečka, J. E., Davis, B. W., Ghosh, S., Paria, N., Das, P. J., Orlando, L., Schubert, M., Nielsen, M. K., Stout, T. A., Brashear, W., & Li, G. (2018). Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nature Communications*, *9*(1), 1–15.

Janssen, A., Colmenares, S., & Karpen, G. (2018). Heterochromatin: Guardian of the Genome. *Annual Review of Cell and Developmental Biology*, *34*, 265–288. https://doi.org/10.1146/ANNUREV-CELLBIO-100617-062653

Janzen, F. J., & Paukstis, G. L. (1991). Environmental sex determination in reptiles: Ecology, evolution, and experimental design. *Quarterly Review of Biology*, *66*(2), 149–179. https://doi.org/10.1086/417143

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., Thomas, M. A., Haussler, D., & Jacob, H. J. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Research*, *4*(14), 528–538. https://doi.org/10.1101/gr.1970304

Jenuwein, T., & David Allis, C. (2001). Translating the Histone Code. *Science*, *293*(5532), 1074–1079.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue), 5–9. https://doi.org/10.1093/nar/gkn201

Kaessmann, H., & Rappold, G. (2018). *The developmental sex-biased expression of genes escaping X chromosome inactivation across mammals*.

Kashimada, K., & Koopman, P. (2010). Sry: The master switch in mammalian sex determination. In *Development* (Vol. 137, Issue 23, pp. 3921–3930). Oxford University Press for The Company of Biologists Limited. https://doi.org/10.1242/dev.048983

Kehr, B., Weese, D., & Reinert, K. (2011). STELLAR: Fast and exact local alignments. *BMC Bioinformatics*, *12*(9), 1–12. https://doi.org/10.1186/1471-2105-12-S9-S15

Kelemen, R. K., & Vicoso, B. (2018). Complex History and Differentiation Patterns of the t-Haplotype, a Mouse Meiotic Driver. *Genetics*, *208*(1), 365–375. https://doi.org/10.1534/genetics.117.300513

Kent, J., Wheatley, S. C., Andrews, J. E., Sinclair, A. H., & Koopman, P. (1996). A male-specific role for SOX9 in vertebrate sex determination. *Development*, *122*(9), 2813–2822.

Kichine, E., Rozé, V., Di Cristofaro, J., Taulier, D., Navarro, A., Streichemberger, E., Decarpentrie, F., Metzler-Guillemain, C., Lévy, N., Chiaroni, J., Paquis-Flucklinger, V., Fellmann, F., & Mitchell, M. J. (2012). HSFY genes and the P4 palindrome in the AZFb interval of the human Y chromosome are not required for spermatocyte maturation. *Human Reproduction*, *27*(2), 615–624. https://doi.org/10.1093/humrep/der421

Kinoshita, K., Shinka, T., Sato, Y., Kurahashi, H., Kowa, H., Chen, G., Umeno, M., Toida, K.,

Kiyokage, E., Nakano, T., Ito, S., & Nakahori, Y. (2006). Expression analysis of a mouse orthologue of HSFY, a candidate for the azoospermic factor on the human Y chromosome. *The Journal of Medical Investigation*, *53*(1,2), 117–122. https://doi.org/10.2152/JMI.53.117

Koepfli, K. P., Paten, B., O'brien, S. J., Antunes, A., Belov, K., Bustamante, C., Castoe, T. A., Clawson, H., Crawford, A. J., Diekhans, M., Distel, D., Durbin, R., Earl, D., Fujita, M. K., Gamble, T., Georges, A., Gemmell, N., Gilbert, M. T. P., Graves, J. M., … Ryder, O. (2015). The genome 10K project: A way forward. *Annual Review of Animal Biosciences*, *3*(1), 57–111. https://doi.org/10.1146/annurev-animal-090414-014900

Koning, A. P. J. de, Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics*, *7*(12), e1002384. https://doi.org/10.1371/JOURNAL.PGEN.1002384

Koopman, P., Ashworth, A., & Lovell-Badge, R. (1991). The ZFY gene family in humans and mice. In *Trends in Genetics* (Vol. 7, Issue 4, pp. 132–136). Elsevier. https://doi.org/10.1016/0168-9525(91)90458-3

Krausz, C., & Casamonti, E. (2017). Spermatogenic failure and the Y chromosome. In *Human Genetics* (Vol. 136, Issue 5, pp. 637–655). Springer Verlag. https://doi.org/10.1007/s00439-017-1793-8

Kruger, A. N., Brogley, M. A., Huizinga, J. L., Kidd, J. M., de Rooij, D. G., Hu, Y. C., & Mueller, J. L. (2019). A Neofunctionalized X-Linked Ampliconic Gene Family Is Essential for Male Fertility and Equal Sex Ratio in Mice. *Current Biology*, *29*(21), 3699-3706.e5. https://doi.org/10.1016/J.CUB.2019.08.057

Lahn, B. T., & Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science*, *286*(5441), 964–967. https://doi.org/10.1126/science.286.5441.964

Lander, S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … Yeh, R.-F. (2001). Initial sequencing and analysis of the human genome. In *Nature* (Vol. 409).

Lange, J., Noordam, M. J., Van Daalen, S. K. M., Skaletsky, H., Clark, B. A., Macville, M. V., Page, D. C., & Repping, S. (2013). Intrachromosomal homologous recombination between inverted amplicons on opposing Y-chromosome arms. *Genomics*, *102*(4), 257–264. https://doi.org/10.1016/j.ygeno.2013.04.018

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Larson, E. L., Kopania, E. E. K., & Good, J. M. (2018). Spermatogenesis and the Evolution of Mammalian Sex Chromosomes. In *Trends in Genetics* (Vol. 34, Issue 9, pp. 722–732). Elsevier Ltd. https://doi.org/10.1016/j.tig.2018.06.003

Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Rg Schibler, J., Tresset, A., Vigne, J.-D., Edwards, C. J., Schlumbaum, A., Dinu, A., Baˇ, A. B., Baˇlaˇbaˇlaˇçsescu, B., Dolman, G., Tagliacozzo, A., Manaseryan, N., Miracle, P., Van Wijngaarden-Bakker, L., Masseti, M., Bradley, D. G., & Cooper, A. (2007). Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Academy of Sciences*, *104*(39), 15276–15281.

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., Lowden, S., Finlayson, H., Brand, T., Willerslev, E., Rowley-Conwy, F., Andersson, L., & Cooper, A. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science*, *307*(5715), 1618–1621. https://doi.org/10.1126/science.1106927

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. ., & Carey, V. J. (2013). R: Finding overlapping genomic ranges. *PLoS Computational Biology*,

9(8), p.e1003118.

Lawrence, Michael, Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), 1003118. https://doi.org/10.1371/journal.pcbi.1003118

Le Rouzic, A., & Capy, P. (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons: Similarities and differences Symbiont influence on host pheromones View project. *Cytogenetic Genome Research*, *110*, 426–440. https://doi.org/10.1159/000084975

Lee, C. T., & Peng, S. L. (2018). A pairwise alignment algorithm for long sequences of high similarity. *Information and Communication Technology*, *625*, 279–287. https://doi.org/10.1007/978-981-10-5508-9_27

Lee, H., & Schatz, M. C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, *28*(16), 2097–2105. https://doi.org/10.1093/bioinformatics/bts330

Li, G., Davis, B. W., Raudsepp, T., Pearks Wilkerson, A. J., Mason, V. C., Ferguson-Smith, M., O, P. C., Waters, P. D., & Murphy, W. J. (2013). Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Research*, *23*(9), 1486–1495. https://doi.org/10.1101/gr.154286.112

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H. L., Nakano, T., & Hotta, A. (2014). Genetic correction using engineered nucleases for gene therapy applications. *Development, Growth & Differentiation*, *56*(1), 63–77.

https://doi.org/10.1111/DGD.12107

Lipman, D. J., Altschul, S. F., & Kececioglut, J. D. (1989). A tool for multiple sequence alignment (proteins/structure/evolution/dynamic programming). *Proceedings of the National Academy of Sciences*, *86*(12), 4412–4415.

Liu, R., Low, W. Y., Tearle, R., Koren, S., Ghurye, J., Rhie, A., Phillippy, A. M., Rosen, B. D., Bickhart, D. M., Smith, T. P. L., Hiendleder, S., & Williams, J. L. (2019). New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC Genomics*, *20*(1), 1–11. https://doi.org/10.1186/s12864-019-6364-z

Livernois, A. M., Graves, J. A. M., & Waters, P. D. (2012). The origin and evolution of vertebrate sex chromosomes and dosage compensation. *Heredity*, *108*(1), 50–58. https://doi.org/10.1038/hdy.2011.106

Lopes. AM, Arnold-Croop, S., Amorim, A., & Carrel, L. (2011). Clustered transcripts that escape X inactivation at mouse XqD. *Mammalian Genome : Official Journal of the International Mammalian Genome Society*, *22*(9–10), 572–582. https://doi.org/10.1007/S00335-011-9350-6

Lucotte, E. A., Skov, L., Jensen, J. M., Macià, M. C., Munch, K., & Schierup, M. H. (2018). Dynamic copy number evolution of X-and Y-linked ampliconic genes in human populations. *Genetics*, *209*(3), 907–920. https://doi.org/10.1534/genetics.118.300826

Lundstörm, K., Matthews, K. R., & Haugen, J. E. (2009). Pig meat quality from entire males. *Animal*, *3*(11), 1497–1507. https://doi.org/10.1017/S1751731109990693

LYON, M. F. (1962). Sex chromatin and gene action in the mammalian X-chromosome. *American Journal of Human Genetics*, *14*(2), 135–148.

Ma, W., Bonora, G., Berletch, J. B., Deng, X., Noble, W. S., & Disteche, C. M. (2018). X-Chromosome inactivation and escape from X inactivation in mouse. In *Methods in Molecular Biology* (Vol. 1861, pp. 205–219). Humana Press Inc. https://doi.org/10.1007/978-1-4939-8766-5_15

Ma, W. J., Veltsos, P., Sermier, R., Parker, D. J., & Perrin, N. (2018). Evolutionary and developmental dynamics of sex-biased gene expression in common frogs with proto-Y chromosomes. *Genome Biology*, *19*(1), 1–17. https://doi.org/10.1186/s13059-018-1548-4

Ma, Y., Zhang, H., Zhang, Q., & Ding, X. (2014). Identification of Selection Footprints on the X Chromosome in Pig. *PLoS ONE*, *9*(4), e94911. https://doi.org/10.1371/journal.pone.0094911

Mácha, J., Teichmanová, R., Sater, A. K., Wells, D. E., Tlapáková, T., Zimmerman, L. B., & Krylov, V. (2012). Deep ancestry of mammalian X chromosome revealed by comparison with the basal tetrapod Xenopus tropicalis. *BMC Genomics*, *13*(1), 315. https://doi.org/10.1186/1471-2164-13-315

MacLean, J. A., Chen, M. A., Wayne, C. M., Bruce, S. R., Rao, M., Meistrich, M. L., Macleod, C., & Wilkinson, M. F. (2005). Rhox: A New Homeobox Gene Cluster. *Cell*, *120*(3), 369–382. https://doi.org/10.1016/J.CELL.2004.12.022

Madden, T. (2003, March 15). *The BLAST Sequence Analysis Tool*. The NCBI Handbook [Internet].; National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK153387/

Mäkinen, V., Salmela, L., & Ylinen, J. (2012). Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics*, *13*(1), 1–5. https://doi.org/10.1186/1471-2105-13-255

Mary, N., Barasc, H., Ferchaud, S., Billon, Y., Meslier, F., Robelin, D., Calgaro, A., Loustau-

Dudez, A.-M., Bonnet, N., Yerle, M., Acloque, H., Ducos, A., & Pinton, A. (2014). Meiotic Recombination Analyses of Individual Chromosomes in Male Domestic Pigs (Sus scrofa domestica). *PLoS ONE*, *9*(6), e99123. https://doi.org/10.1371/journal.pone.0099123

Matsuno, Y., Yamashita, T., Wagatsuma, M., & Yamakage, H. (2019). Convergence in LINE-1 nucleotide variations can benefit redundantly forming triplexes with lncRNA in mammalian X-chromosome inactivation. *Mobile DNA*, *10*(1), 33. https://doi.org/10.1186/s13100-019-0173-4

Mccoard, S. A., Fahrenkrug, S. C., Alexander, L. J., Freking, B. A., Rohrer, G. A., Wise, T. H., & Ford, J. J. (2002). An integrated comparative map of the porcine X chromosome. *Animal Genetics* , *33*, 178–185.

Megens, H. J., & Groenen, M. A. M. (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. In *Heredity* (Vol. 109, Issue 1, pp. 1–3). Nature Publishing Group. https://doi.org/10.1038/hdy.2011.128

Mercer, T. R., Dinger, M. E., & Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* , *10*(3), 155–159. https://doi.org/10.1038/nrg2521

Miklenic, M., & Svetec, I. (2021). Palindromes in DNA—A Risk for Genome Stability and Implications in Cancer. *International Journal of Molecular Sciences*, *22*(2840).

Mills, R., Bennett, E., Iskow, R., Luttig, C., Tsui, C., Pittard, W., & Devine, S. (2006). Recently mobilized transposons in the human and chimpanzee genomes. *American Journal of Human Genetics*, *78*(4), 671–679. https://doi.org/10.1086/501028

Molaro, A., Wood, A. J., Janssens, D., Kindelay, S. M., Eickbush, M. T., Wu, S., Singh, P., Muller, C. H., Henikoff, S., & Malik, H. S. (2020). Biparental contributions of the H2A.B histone

variant control embryonic development in mice. *PLOS Biology*, *18*(12), e3001001. https://doi.org/10.1371/JOURNAL.PBIO.3001001

Molaro, A., Young, J. M., & Malik, H. S. (2018). Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals. *Genome Research*, *28*(4), 460–473. https://doi.org/10.1101/GR.229799.117

Moretti, C., Blanco, M., Ialy-Radio, C., Serrentino, M.-E., Gobé, C., Friedman, R., Battail, C., Leduc, M., Ward, M. A., Vaiman, D., Tores, F., & Cocquet, J. (2020). Battle of the Sex Chromosomes: Competition between X and Y Chromosome-Encoded Proteins for Partner Interaction and Chromatin Occupancy Drives Multicopy Gene Expression and Evolution in Muroid Rodents. *Molecular Biology and Evolution*, *37*(12), 3453–3468. https://doi.org/10.1093/molbev/msaa175

Moretti, C., Vaiman, D., Tores, F., & Cocquet, J. (2016). Expression and epigenomic landscape of the sex chromosomes in mouse post-meiotic male germ cells. *Epigenetics & Chromatin*, *9*(1), 1–18. https://doi.org/10.1186/s13072-016-0099-8

Mount, D. (2003). *Bioinformatics: Sequence and Genome Analysis.* (2nd ed.). Cold Spring Harbour Laboratory Press: Cold Spring Harbour.

Mueller, J. L., Mahadevaiah, S. K., Park, P. J., Warburton, P. E., Page, D. C., & Turner, J. M. A. (2008). The mouse X chromosome is enriched for multi-copy testis genes exhibiting post-meiotic expression. *Nature Genetics*, *40*(6), 794. https://doi.org/10.1038/NG.126

Mueller, J. L., Skaletsky, H., Brown, L. G., Zaghlul, S., Rock, S., Graves, T., Auger, K., Warren, W. C., Wilson, R. K., & Page, D. C. (2013). Independent specialization of the human and mouse X chromosomes for the male germ line. *Nature Genetics*, *45*(9), 1083–1087. https://doi.org/https://doi.org/10.1038/ng.2705

Muller, H. . (1918). Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics*, *3*(5), 422.

Muller, H. . (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *1*(1), 2–9.

Murakami, H., & Keeney, S. (2008). Regulating the formation of DNA double-strand breaks in meiosis. In *Genes and Development* (Vol. 22, Issue 3, pp. 286–292). Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gad.1642308

Musio, A. (2020). The multiple facets of the SMC1A gene. *Gene*, *743*, 144612. https://doi.org/10.1016/J.GENE.2020.144612

Naidu, S. D., & Dinkova-Kostova, A. T. (2017). Regulation of the mammalian heat shock factor 1. *The FEBS Journal*, *284*(11), 1606–1627. https://doi.org/10.1111/FEBS.13999

Nailwal, M., & Chauhan, J. B. (2017). Computational analysis of high risk missense variant in human UTY Gene: A candidate gene of AZFa sub-region. *Journal of Reproduction and Infertility*, *18*(3), 298–306.

Nam, K., Munch, K., Hobolth, A., Dutheil, J. Y., Veeramah, K. R., Woerner, A. E., Hammer, M. F., Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., … Schierup, M. H. (2015). Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(20), 6413–6418. https://doi.org/10.1073/pnas.1419306112

Ngamphiw, C., Tongsima, S., & Mutirangura, A. (2014). Roles of Intragenic and Intergenic L1s in Mouse and Human. *PLOS ONE*, *9*(11), e113434. https://doi.org/10.1371/JOURNAL.PONE.0113434

Nguyen, D., & Disteche, C. (2006). Dosage compensation of the active X chromosome in mammals. *Nature Genetics*, *38*(1), 47–53.

Nguyen, D. T., Lee, K., Choi, H., Choi, M. kyeung, Le, M. T., Song, N., Kim, J. H., Seo, H. G., Oh, J. W., Lee, K., Kim, T. H., & Park, C. (2012). The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. *BMC Genomics*, *13*(1), 1–12. https://doi.org/10.1186/1471-2164-13-584

Nguyen, H. T., Johnson, A. F., Nguyen, H. T., & Veitia, R. A. (2019). Causes and effects of haploinsufficiency. *Biological Reviews*, *94*(5), 1774–1785. https://doi.org/10.1111/brv.12527

Ohno, S. (1966). Conservation of the Original X and Homology of the X-linked Genes in Placental Mammals. In *In Sex Chromosomes and Sex-linked Genes* (pp. 46–73). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-35113-0_5

Ohno, S. (2013). *Sex Chromosomes and Sex-Linked Genes* (Vol. 1). Springer Science & Business Media .

Olender, T., Lancet, D., & Nebert, D. W. (2008). Update on the olfactory receptor (OR) gene superfamily. In *Human genomics* (Vol. 3, Issue 1, pp. 1–11). BioMed Central. https://doi.org/10.1186/1479-7364-3-1-87

Oluwole, O. A., Mahboubi, K., Favetta, L. A., Revay, T., Kroetsch, T., & King, W. A. (2017). Highly dynamic temporal changes of TSPY gene copy number in aging bulls. *PLOS ONE*, *12*(5), e0178558. https://doi.org/10.1371/JOURNAL.PONE.0178558

Pairwise Sequence Alignment Tools < EMBL-EBI. (n.d.). European Nucleotide Archive. Retrieved 4 January 2021, from https://www.ebi.ac.uk/Tools/psa/

Panning, B., & Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and

silence X-linked genes. *Genes and Development*, *10*(16), 1991–2002.

Paria, N., Raudsepp, T., Pearks Wilkerson, A. J., O'Brien, P. C. M., Ferguson-Smith, M. A., Love, C. C., Arnold, C., Rakestraw, P., Murphy, W. J., & Chowdhary, B. P. (2011). A Gene Catalogue of the Euchromatic Male-Specific Region of the Horse Y Chromosome: Comparison with Human and Other Mammals. *PLoS ONE*, *6*(7), e21374. https://doi.org/10.1371/journal.pone.0021374

Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A. F., Bosse, M., Bastiaansen, J. W. M., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics*, *14*(1), 449. https://doi.org/10.1186/1471-2164-14-449

Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A. F., Bosse, M., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2015). Copy number variation in the speciation of pigs: A possible prominent role for olfactory receptors. *BMC Genomics*, *16*(1), 1–14. https://doi.org/10.1186/s12864-015-1449-9

Pig [Sus scrofa] Genomic Dataset. (n.d.). Retrieved 2 October 2021, from http://repeatmasker.org/species/susScr.html

Pirkkala, L., Nykänen, P., & Sistonen, L. E. A. (2001). Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. *The FASEB Journal*, *15*(7), 1118–1131. https://doi.org/10.1096/fj00-0294rev

Pontier, D. B., & Gribnau, J. (2011). Xist regulation and function eXplored. *Human Genetics 2011 130:2*, *130*(2), 223–236. https://doi.org/10.1007/S00439-011-1008-7

Popova, B. C., Tada, T., Takagi, N., Brockdorff, N., & Nesterova, T. B. (2006). Attenuated spread of X-inactivation in an X;autosome translocation. *Proceedings of the National Academy of*

Sciences of the United States of America, *103*(20), 7706–7711. https://doi.org/10.1073/pnas.0602021103

Prather, R. S., Shen, M., & Dai, Y. (2008). Genetically Modified Pigs for Medicine and Agriculture. *Biotechnology and Genetic Engineering Reviews*, *25*(1), 245–266. https://doi.org/10.5661/bger-25-245

Rabindran, S., Giorgi, G., Clos, J., & Wu, C. (1991). Molecular cloning and expression of a human heat shock factor, HSF1 (human transcription factor/leudne zippers/polymerase chain reaction). *Proceedings of the National Academy of Sciences*, *88*(16), 6906–6910.

Ramathal, C., Angulo, B., Sukhwani, M., Cui, J., Durruthy-Durruthy, J., Fang, F., Schanes, P., Turek, P. J., Orwig, K. E., & Reijo Pera, R. (2015). DDX3Y gene rescue of a y chromosome AZFa deletion restores germ cell formation and transcriptional programs. *Scientific Reports*, *5*(1), 1–13. https://doi.org/10.1038/srep15041

Rastan, S., & Robertson, E. J. (1985). X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation. In *J. Embryol. exp. Morph* (Vol. 90).

Rathje, C. C., Johnson, E. E. P., Drage, D., Patinioti, C., Silvestri, G., Affara, N. A., Ialy-Radio, C., Cocquet, J., Skinner, B. M., & Ellis, P. J. I. (2019). Differential Sperm Motility Mediates the Sex Ratio Drive Shaping Mouse Sex Chromosome Evolution. *Current Biology*, *29*(21), 3692-3698.e4. https://doi.org/10.1016/j.cub.2019.09.031

Raudsepp, T., Das, P. J., Avila, F., & Chowdhary, B. P. (2012). The Pseudoautosomal Region and Sex Chromosome Aneuploidies in Domestic Species. *Sexual Development*, *6*(1–3), 72–83. https://doi.org/10.1159/000330627

Raudsepp, T., Kata, S. R., Piumi, F., Swinburne, J., Womack, J. E., Skow, L. C., & Chowdhary, B. P. (2002). Conservation of gene order between horse and human X chromosomes as

evidenced through radiation hybrid mapping. *Genomics*, *3*(79), 451–457.

Raudsepp, Terje, & Chowdhary, B. P. (2015). The eutherian pseudoautosomal region. In *Cytogenetic and Genome Research* (Vol. 147, Issues 2–3, pp. 81–94). S. Karger AG. https://doi.org/10.1159/000443157

Rauschendorf, M.-A., Zimmer, J., Ohnmacht, C., & Vogt, P. H. (2014). DDX3X, the X homologue of AZFa gene DDX3Y, expresses a complex pattern of transcript variants only in the male germ line. *Molecular Human Reproduction*, *20*(12), 1208–1222. https://doi.org/10.1093/molehr/gau081

Rautiala, P., Helanterä, H., & Puurtinen, M. (2019). Extended haplodiploidy hypothesis. *Evolution Letters*, *3*(3), 263–270. https://doi.org/10.1002/evl3.119

Repping, S., Skaletsky, H., Brown, L., M van Daalen, S. K., Korver, C. M., Pyntikova, T., Kuroda-Kawaguchi, T., A de Vries, J. W., Oates, R. D., Silber, S., van der Veen, F., Page, D. C., & Rozen, S. (2003). Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nature Genetics*, *35*(3), 247–251. https://doi.org/10.1038/ng1250

Rice, W. R. (1987). The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes. In *Evolution* (Vol. 41, Issue 4).

Rice, W. R. (1996). Evolution of the Y sex chromosome in animals. *BioScience*, *46*(5), 331–343. https://doi.org/10.2307/1312947

Rice, W. R. (2013). Nothing in Genetics Makes Sense Except in Light of Genomic Conflict. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 217–237. https://doi.org/10.1146/annurev-ecolsys-110411-160242

Rodriguez-Terrones, D., & Torres-Padilla, M.-E. (2018). Nimble and Ready to Mingle: Transposon Outbursts of Early Development Box 1. Transposable Elements Exhibit Limited Conservation across Mammals. *Trends in Genetics*, *34*(10), 806–820. https://doi.org/10.1016/j.tig.2018.06.006

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., Prankish, A., Lovell, F. L., Howe, K. L., Ashurst, J. L., Fulton, R. S., Sudbrak, R., Wen, G., Jones, M. C., Hurles, M. E., … Bentley, D. R. (2005). The DNA sequence of the human X chromosome. *Nature*, *434*(7031), 325–337. https://doi.org/10.1038/nature03440

Rothschild, M. F., & Ruvinsky, A. (2011). *The Genetics of the Pig* (Second edi). CABI.

Roze, D., & Barton, N. H. (2006). The Hill-Robertson effect and the evolution of recombination. *Genetics*, *173*(3), 1793–1811. https://doi.org/10.1534/genetics.106.058586

Rubin, C.-J., Megens, H.-J., Martinez Barrio, A., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C. B., Archibald, A. L., Fredholm, M., Groenen, M. A. M., & Andersson, L. (2012). Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences*, *109*(48), 19529–19536. https://doi.org/10.1073/pnas.1217149109

Sahakyan, A., Plath, K., & Rougeulle, C. (2017). Regulation of X-chromosome dosage compensation in human: mechanisms and model systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1733), 20160363. https://doi.org/10.1098/rstb.2016.0363

Sandstedt, S. A., & Tucker, P. K. (2004). Evolutionary Strata on the Mouse X Chromosome Correspond to Strata on the Human X Chromosome. *Genome Research*, *14*(2), 267–272. https://doi.org/10.1101/gr.1796204

Sato, Y., Yoshida, K., Shinka, T., Nozawa, S., Nakahori, Y., & Iwamoto, T. (2006). Altered expression pattern of heat shock transcription factor, Y chromosome (HSFY) may be related to altered differentiation of spermatogenic cells in testes with deteriorated spermatogenesis. *Fertility and Sterility*, *86*(3), 612–618. https://doi.org/10.1016/j.fertnstert.2006.01.053

Saxena, R. (2000). *The human Deleted in Azoospermia gene family : structure, function and evolution*. Massachusetts Institute of Technology.

Schook, L. B., Beever, J. E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., Milan, D., Rohrer, G., & Eversole, K. (2005). Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. *Comparative and Functional Genomics*, *6*(4), 251–255. https://doi.org/10.1002/cfg.479

Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and complex phenotypes. In *Current Biology* (Vol. 24, Issue 7, pp. R288–R294). Cell Press. https://doi.org/10.1016/j.cub.2014.01.056

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., & Miller, W. (2003). Human-Mouse Alignments with BLASTZ. *Genome Research*, *13*(1), 103–107. https://doi.org/10.1101/gr.809403

Schwarz, D. G. (2000). Related function of mouse SOX3, SOX9, and SRY HMG domains assayed by male sex determination. *Genesis*, *28*(3–4), 111–124. https://doi.org/10.1002/1526-968X(200011/12)28:3/4<111::AID-GENE40>3.0.CO;2-5

Scott, T., & West, S. (2019). Adaptation is maintained by the parliament of genes. *Nature Communications*, *10*(1), 1–13.

Seefelder, M., Alva, V., Huang, B., Engler, T., Baumeister, W., Guo, Q., Fernández-Busnadiego, R., Lupas, A. N., & Kochanek, S. (2020). The evolution of the huntingtin-associated protein

40 (HAP40) in conjunction with huntingtin. *BMC Evolutionary Biology* , *20*(1), 1–18. https://doi.org/10.1186/S12862-020-01705-5

Shu, J.-J., & Shan Ouw, L. (2004). Pairwise alignment of the DNA sequence using hypercomplex number representation. *Bulletin of Mathematical Biology*, *66*(5), 1423–1438. https://doi.org/10.1016/j.bulm.2004.01.005

Silver, L. M. (1989). Gene dosage effects on transmission ratio distortion and fertility in mice that carry t haplotypes. *Genetics Research*, *54*(3), 221–225. https://doi.org/10.1017/S0016672300028688

Simon, M. D., Pinter, S. F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S. K., Kesner, B. A., Maier, V. K., Kingston, R. E., & Lee, J. T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, *504*(7480), 465–469. https://doi.org/10.1038/nature12719

Skaletsky, H., Kuroda-Kawaguchl, T., Minx, P. J., Cordum, H. S., Hlllier, L. D., Brown, L. G., Repplng, S., Pyntikova, T., All, J., Blerl, T., Chinwalla, A., Delehaunty, A., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S. F., Latrielle, P., … Page, D. C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*(6942), 825–837. https://doi.org/10.1038/nature01722

Skinner, B. M., Lachani, K., Sargent, C. A., & Affara, N. A. (2013). Regions of XY homology in the pig X chromosome and the boundary of the pseudoautosomal region. *BMC Genomics*, *14*(1), 1–7. https://doi.org/10.1186/1471-2156-14-3

Skinner, B. M., Lachani, K., Sargent, C. A., Yang, F., Ellis, P., Hunt, T., Fu, B., Louzada, S., Churcher, C., Tyler-Smith, C., & Affara, N. A. (2015). Expansion of the HSFY gene family in pig lineages. *BMC Genomics*, *16*(1), 1–11. https://doi.org/10.1186/s12864-015-1650-x

Skinner, B. M., Sargent, C. A., Churcher, C., Hunt, T., Herrero, J., Loveland, J. E., Dunn, M., Louzada, S., Fu, B., Chow, W., Gilbert, J., Austin-Guest, S., Beal, K., Carvalho-Silva, D., Cheng, W., Gordon, D., Grafham, D., Hardy, M., Harley, J., … Tyler-Smith, C. (2016). The pig X and Y Chromosomes: structure, sequence, and evolution. *Genome Research*, *26*(1), 130–139. https://doi.org/10.1101/gr.188839.114

Skuse, D. H. (2005). X-linked genes and mental functioning. In *Human Molecular Genetics* (Vol. 14, Issue SPEC. ISS. 1, pp. 27–32). Oxford Academic. https://doi.org/10.1093/hmg/ddi112

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. In *Nature Reviews Genetics* (Vol. 8, Issue 4, pp. 272–285). Nature Publishing Group. https://doi.org/10.1038/nrg2072

Small, K., Iber, J., & Warren, S. T. (1997). Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nature Genetics* , *16*(1), 96–99. https://doi.org/10.1038/ng0597-96

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart - Biological queries made easy. *BMC Genomics*, *10*(1), 1–12. https://doi.org/10.1186/1471-2164-10-22

Smeds, L., Kojola, I., & Ellegren, H. (2019). The evolutionary history of grey wolf Y chromosomes. *Molecular Ecology*, *28*(9), 2173–2191. https://doi.org/10.1111/mec.15054

Smit, A. F. ., Hubley, R., & Green, P. (n.d.). *RepeatMasker Home Page*. ISB. Retrieved 21 November 2020, from http://www.repeatmasker.org/

Smit, A. F. ., Hubley, R., & Green, P. (1999). *RepeatMasker Documentation*. ISB. http://www.repeatmasker.org/webrepeatmaskerhelp.html

Smith, C. A., Roeszler, K. N., Ohnesorg, T., Cummins, D. M., Farlie, P. G., Doran, T. J., &

Sinclair, A. H. (2009). The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature*, *461*(7261), 267–271. https://doi.org/10.1038/nature08298

Soh, Y. Q. S., Alföldi, J., Pyntikova, T., Brown, L. G., Graves, T., Minx, P. J., Fulton, R. S., Kremitzki, C., Koutseva, N., Mueller, J. L., Rozen, S., Hughes, J. F., Owens, E., Womack, J. E., Murphy, W. J., Cao, Q., De Jong, P., Warren, W. C., Wilson, R. K., … Page, D. C. (2014). Sequencing the mouse y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell*, *159*(4), 800–813. https://doi.org/10.1016/j.cell.2014.09.052

Song, H. W., Anderson, R. A., Bayne, R. A., Gromoll, J., Shimasaki, S., Chang, R. J., Parast, M. M., Laurent, L. C., de Rooij, D. G., Hsieh, T. C., & Wilkinson, M. F. (2013). The RHOX homeobox gene cluster is selectively expressed in human oocytes and male germ cells. *Human Reproduction*, *28*(6), 1635–1646. https://doi.org/10.1093/HUMREP/DET043

Sonnhammer, E. L. L., & Durbin, R. (1996). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. In *Gene* (Vol. 167).

Sousa, L., Barros De, Andrade EJonkers, I., Syx, L., Dunkel, I., Chaumeil, J., Picard, C., Foret, B., Chen, C.-J., Lis, J. T., Heard, E., Schulz, E. G., & Marsico, A. (2019). Kinetics of Xist-induced gene silencing can be predicted from combinations of epigenetic and genomic features. *Genome Research*, *29*(7), 1087–1099. https://doi.org/10.1101/gr.245027.118

Statello, L., Guo, C.-J., Chen, L.-L., & Huarte, M. (2020). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology* , *22*(2), 96–118. https://doi.org/10.1038/s41580-020-00315-9

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Project, 1000 Genomes, & Eichler, E. E. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. *Science (New York, N.Y.)*, *330*(6004), 641.

https://doi.org/10.1126/SCIENCE.1197005

Sugimoto, M. (2014). Developmental genetics of the mouse t-complex. In *Genes & Genetic systems* (Vol. 89).

Szak, S., Pickeral, O., Makalowski, W., Boguski, M., Landsman, D., & Boeke, J. (2002). Molecular archeology of L1 insertions in the human genome. *Genome Biology*, *3*(10). https://doi.org/10.1186/GB-2002-3-10-RESEARCH0052

Talbert, P. B., & Henikoff, S. (2010). Centromeres Convert but Don't Cross. *PLoS Biology*, *8*(3). https://doi.org/10.1371/JOURNAL.PBIO.1000326

Tang, Y. A., Huntley, D., Montana, G., Cerase, A., Nesterova, T. B., & Brockdorff, N. (2010). Efficiency of Xist-mediated silencing on autosomes is linked to chromosomal domain organisation. *Epigenetics & Chromatin 2010 3:1*, *3*(1), 1–12. https://doi.org/10.1186/1756-8935-3-10

Tarailo-Graovac, M., & Chen, N. (2004). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, *25*(1), 4–10. https://doi.org/10.1002/0471250953.bi0410s25

Teixeira, S. A., Ibelli, A. M. G., Cantão, M. E., Oliveira, H. C. de, Ledur, M. C., Peixoto, J. de O., Marques, D. B. D., Costa, K. A., Coutinho, L. L., & Guimarães, S. E. F. (2019). Sex Determination Using RNA-Sequencing Analyses in Early Prenatal Pig Development. *Genes*, *10*(12), 1010. https://doi.org/10.3390/genes10121010

Tempel, S. (2012). Using and Understanding RepeatMasker. *Methods in Molecular Biology*, *859*, 29–51. https://doi.org/10.1007/978-1-61779-603-6_2

Tessari, A., Ferlin, A., Salata, A., Bartoloni, L., Slongo, M., & Foresta, C. (2004). Characterization of HSFY, a novel AZFb gene on the Y chromosome with a possible role in human

spermatogenesis. *Molecular Human Reproduction*, *10*(4), 253–258. https://doi.org/10.1093/molehr/gah036

Tolstorukov, M., Goldman, J., Gilbert, C., Ogryzko, V., Kingston, R., & Park, P. (2012). Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Molecular Cell*, *47*(4), 596–607. https://doi.org/10.1016/J.MOLCEL.2012.06.011

Touré, A., Szot, M., Mahadevaiah, S. K., Ine Rattigan, A. ′, Ojarikre, O. A., & Burgoyne, P. S. (2004). A New Deletion of the Mouse Y Chromosome Long Arm Associated With the Loss of Ssty Expression, Abnormal Sperm Development and Sterility. *Genetics*, *166*(2), 901–912.

Treangen, T. J., & Messeguer, X. (2006). M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics*, *7*(1), 1–15. https://doi.org/10.1186/1471-2105-7-433

Tukiainen, T., Villani, A., & Yen, A. (2017). Landscape of X chromosome inactivation across human tissues. *Nature*, *550*, 244–248.

Vallender, E. J., & Lahn, B. T. (2004). How mammalian sex chromosomes acquired their peculiar gene content. *BioEssays*, *26*, 159–169. https://doi.org/10.1002/bies.10393

Vegesna, R., Tomaszkiewicz, M., Medvedev, P., & Makova, K. D. (2019). Dosage regulation, and variation in gene expression and copy number of human Y chromosome ampliconic genes. *PLoS Genetics*, *15*(9), e1008369. https://doi.org/10.1371/journal.pgen.1008369

Vockel, M., Riera-Escamilla, A., Tüttelmann, F., & Krausz, C. (2019). The X chromosome and male infertility. *Human Genetics*, *140*(1), 203–215. https://doi.org/10.1007/S00439-019-02101-W

Vodicka, R., Vrtel, R., Dusek, L., Singh, A. R., Krizova, K., Svacinova, V., Horinova, V., Dostal, J., Oborna, I., Brezinova, J., Sobek, A., & Santavy, J. (2007). TSPY gene copy number as a

potential new risk factor for male infertility. *Reproductive Biomedicine Online*, *14*(5), 579–587. https://doi.org/10.1016/S1472-6483(10)61049-8

Wainer Katsir, K., & Linial, M. (2019). Human genes escaping X-inactivation revealed by single cell expression data. *BMC Genomics*, *20*(1), 1–17. https://doi.org/10.1186/s12864-019-5507-6

Wang, J., Jiang, J., Wang, H., Kang, H., Zhang, Q., & Liu, J.-F. (2014). Enhancing Genome-Wide Copy Number Variation Identification by High Density Array CGH Using Diverse Resources of Pig Breeds. *PLOS ONE*, *9*(1), e87571. https://doi.org/10.1371/JOURNAL.PONE.0087571

Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y., & Benson, G. (2004). Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeats That Contain Testes Genes. *Genome Research*, *14*(10A), 1861–1869. https://doi.org/10.1101/gr.2542904

Ward, M. A., & Burgoyne, P. S. (2006). The effects of deletions of the mouse Y chromosome long arm on sperm function - Intracytoplasmic sperm injection (ICSI)-based analysis. *Biology of Reproduction*, *74*(4), 652–658. https://doi.org/10.1095/biolreprod.105.048090

Warner, D. A., & Shine, R. (2008). The adaptive significance of temperature-dependent sex determination in a reptile. *Nature*, *451*(7178), 566–568. https://doi.org/10.1038/nature06519

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., Chow, W., Eory, L., Finlayson, H. A., Flicek, P., Girón, C. G., Griffin, D. K., Hall, R., Hannum, G., Hourlier, T., Howe, K., Hume, D. A., Izuogu, O., Kim, K., … Archibald, A. L. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience*, *9*(6), 1–14. https://doi.org/10.1093/gigascience/giaa051

Webster, T. H., Couse, M., Grande, B. M., Karlins, E., Phung, T. N., Richmond, P. A., Whitford,

W., & Wilson, M. A. (2019). Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience*, *8*(7), 1–11. https://doi.org/10.1093/gigascience/giz074

Werren, J. H. (1987). The Coevolution of Autosomal and Cytoplasmic Sex Ratio Factors. *Journal of Theoretical Biology*, *124*(3), 317–334.

West, S. (2009). *Sex Allocation* (Vol. 44).

Whyte, J. J., & Prather, R. S. (2011). Genetic modifications of pigs for medicine and agriculture. *Molecular Reproduction and Development*, *78*(10–11), 879–891. https://doi.org/10.1002/mrd.21333

Wichman, H. A., Bussche, R. A. Van Den, Hamilton, M. J., & Baker, R. J. (1993). Transposable elements and the evolution of genome organization in mammals. *Transposable Elements and Evolution. Contemporary Issues in Genetics and Evolution* , *1*, 149–157. https://doi.org/10.1007/978-94-011-2028-9_11

Widlak, W., & Vydra, N. (2017). The Role of Heat Shock Factors in Mammalian Spermatogenesis. *Advances in Anatomy Embryology and Cell Biology*, *222*, 45–65. https://doi.org/10.1007/978-3-319-51409-3_3

William Roy, S., & Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* , *7*(3), 211–221. https://doi.org/10.1038/nrg1807

Wright, A., Zimmer, D., & Mank, J. E. (2016). How to make a sex chromosome. *Nature Communications*, *7*(1), 1–8.

Wright, S. I., Kent, T. V, & Uzunovic´, J. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biologival Sciences*, *372*(1736). https://doi.org/10.1098/rstb.2016.0458

Wu, C. (1995). Heat shock transcription factors: structure and regulation. *Annual Review of Cell and Developmental Biology*, *11*, 441–469. https://doi.org/10.1146/ANNUREV.CB.11.110195.002301

Yatskevich, S., Rhodes, J., & Nasmyth, K. (2019). Organization of Chromosomal DNA by SMC Complexes. *Annual Review of Genetics*, *53*, 445–482. https://doi.org/10.1146/ANNUREV-GENET-112618-043633

Yazdi, H. P., Silva, W. T. A. F., & Suh, A. (2020). Why do some sex chromosomes degenerate more slowly than others? The odd case of ratite sex chromosomes. *Genes*, *11*(10), 1–13. https://doi.org/10.3390/genes11101153

Ye, D., Zaidi, A. A., Tomaszkiewicz, M., Anthony, K., Liebowitz, C., DeGiorgio, M., Shriver, M. D., & Makova, K. D. (2018). High levels of copy number variation of ampliconic genes across major human Y haplogroups. *Genome Biology and Evolution*, *10*(5), 1333–1350. https://doi.org/10.1093/gbe/evy086

Yin, T., Cook, D., & Lawrence, M. (2012). ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biology*, *13*(8), 1–14. https://doi.org/10.1186/gb-2012-13-8-r77

Yue, X. P., Dechow, C., Chang, T. C., DeJarnette, J. M., Marshall, C. E., Lei, C. Z., & Liu, W. S. (2014). Copy number variations of the extensively amplified Y-linked genes, HSFY and ZNF280BY, in cattle and their association with male reproductive traits in Holstein bulls. *BMC Genomics*, *15*(1), 1–12. https://doi.org/10.1186/1471-2164-15-113

Zafar, F., Okita, A. K., Onaka, A. T., Su, J., Katahira, Y., Nakayama, J., Takahashi, T. S., Masukata, H., & Nakagawa, T. (2017). Regulation of mitotic recombination between DNA repeats in centromeres. *Nucleic Acids Research*, *45*(19), 11222. https://doi.org/10.1093/NAR/GKX763

Zanders, S. E., & Unckless, R. L. (2019). Fertility Costs of Meiotic Drivers. In *Current Biology* (Vol. 29, Issue 11, pp. R512–R520). Cell Press. https://doi.org/10.1016/j.cub.2019.03.046

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, W., Yang, M., Zhou, M., Wang, Y., Wu, X., Zhang, X., Ding, Y., Zhao, G., Yin, Z., & Wang, C. (2020). Identification of Signatures of Selection by Whole-Genome Resequencing of a Chinese Native Pig. *Frontiers in Genetics*, *11*, 566255. https://doi.org/10.3389/fgene.2020.566255

Zhang, X., & Firestein, S. (2002). The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience*, *5*(2), 124–133. https://doi.org/10.1038/nn800

Zhang, Y. E., Vibranovski, M. D., Krinsky, B. H., & Long, M. (2011). A cautionary note for retrocopy identification: DNA-based duplication of intron-containing genes significantly contributes to the origination of single exon genes. *Bioinformatics*, *27*(13), 1749–1753. https://doi.org/10.1093/BIOINFORMATICS/BTR280

Zhang, Yang. (n.d.). *FASTA format*. Zhang Lab. Retrieved 29 June 2021, from https://zhanglab.dcmb.med.umich.edu/FASTA/

Zhang, Yi, & Reinberg, D. (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes & Development*, *15*(8), 2343–2360. https://doi.org/10.1101/gad.927301