Investigation of the genetic components of maternal infanticide in *Sus scrofa*

Julien Bauer

A thesis presented for the degree of

Doctor of Philosophy

September 2018

Hughes Hall

Abstract

Investigation of the genetic components of maternal infanticide in *Sus scrofa*

Julien Bauer

The aim of this thesis was to investigate the genetic components of maternal infanticide in the domestic pig (*Sus scrofa*). The killing of piglets by sows by eating them soon after birth is a significant issue for the pig breeding industry because of the impact on animal welfare and the cost in lost revenues. There is evidence that part of this behavior has a genetic basis and this work aims at finding the genes and genome regions linked to this trait. This study focuses on animals from four different breed lines; two have around a 5% incidence of infanticide and two have around a 10% incidence of infanticide.

The first half of this work used a genotyping approach using the pig 60K SNP array from Illumina. Two different tests were used to analyse the data: Family Based Association Test (FBAT) and Parent of Origin (PO) test. The FBAT approach uses pedigree information to test for association in the presence of linkage and the PO approach tests for preferential transmission of allele from one parent (this study focused on maternal transmission).

The results from the tests, along with previous results generated by our group, were used to design three sequence capture sets in order to study these regions in more detail. The sequencing work was done on a selection of animals grouped in pools, for each line two pools of infanticide animals were selected: animals with an history of infanticide in the pedigree and animals with multiple instances of infanticide. Once sequencing was completed, variants were called on the region of interest, for each pool and the different capture sets, using the Genome Analysis Tool Kit. The variant allele frequencies in the pool was compared between control pools and infanticide pools to select target variants. Some of the variants identified are interesting targets and identify genes of interests.

Declarations

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Collaborations

- Dr Claire Quilter for the array work and the design of the sequence capture sets
- Miss Kerry Harvey for the processing of the array and construction of the sequence capture sets
- The team at Cambridge Genomic Services for the sequencing of the libraries

Acknowledgements

This thesis has been present in my life for a long time. If I had a baby when I started, it would be seven years old now! It has been an interesting journey and a challenging one. Working on this thesis while having a full time job proved quite a challenge. Thankfully, some great people supported me along the way. I would like to start by thanking Professor Nabeel Affara who made this work possible and helped me getting started on this part time adventure. This work would not have seen completion without the help of Dr. Carole Sargent and Dr. Claire Quilter. Their advice and knowledge on the subject have proven invaluable. I would like also to thank the people in the Affara lab for their help and support, in particular the Cambridge Genomic Services team. They have been patient and took a lot on their shoulders to help me complete this thesis. Finally I want to thank my wife, Sandra Bauer, who pushed me to take on this project and helped keeping me motivated. During those seven years I learned a lot of news skills on how to design pipelines and scripts in order to process large amounts of data quickly. Now that it is finished, I will have more time to dedicate to my hobby of painting miniatures, and playing board games with friends and members of the lab!

Abbreviations

General abbreviations

A	adenine
AA	Amino Acid
ADHD	attention deficit hyperactivity disorder
Bp	base pair
BWA	Burrow Wheeler Aligner
BWT	Burrow Wheeler Transform
C	cytosine
CGH	Comparative Genomic Hybridisation
ChIP	Chromatin Immuno Precipitation
DNA	Deoxyribonucleic acid
FBAT	Family Based Association Test
G	guanine
GATK	Genome Analysis Tool Kit
GRR	Genotypic Risk Ratio
GWAS	Genome Wide Association Study
HG	Human Genome
HMM	Hidden Markov Models
IBD	Identity By Descent
IBS	Identity By State
kb	kilo bases
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MB	Mega Base

OR	Odd Ratio
PCR	Polymerase Chain Reaction
PO	Parent of Origin
QTL	Quantitative Trait Loci
RNA	ribonucleic acid
RNASeq	RNA sequencing
SNP	Single Nucleotide Polymorphism
SSC	*Sus scrofa* Chromosome
SMRT	Single-Molecule Real-Time
SS	*Sus scrofa*
T	thymine
TDT	Transmission Disequilibrium Test
X	folds

Genes abbreviation

5HTR2C	5-hydroxytryptamine (serotonin) receptor 2C
ABCB1	ATP binding cassette subfamily B member 1
ADORA2A	adenosine A2a receptor
ADRA2A	Alpah-2A adrenergic receptor
APSM	abnormal spindle microtubule assembly
ARPP21	cAMP regulated phosphoprotein 21
ASAP1	ArfGAP with SH3 domain
AUTS2	AUTS2
AVP	arginine vasopressin
BMPER	BMP binding endothelial regulator
CDH11	cadherin 11
CLC1	chloride voltage-gated channel 1
CLCC1	chloride channel CLIC like 1
CLIC6	chloride intracellular channel 6
COL4A3BP	collagen type IV alpha 3 binding protein [Homo sapiens (human)]
COX7C	cytochrome subunit VIIc
CYMP	chymosin pseudogene
DPB	Albumin D-site-binding protein
DRD2	dopamine receptor 2
EXT1	exostosin glycosyltransferase 1
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit
FAM114A2	family with sequence similarity 114 member A2
FRMD4A	FERM domain containing 4A
GLRA2	Glycine Receptor Alpha2
GNAI1	Adenylase cyclase unhibiting G alpha protein
GPM6A	glycoprotein M6A
GPR139	G protein-coupled receptor 139
GPR158	G protein-coupled receptor 158
GPSM2	(G protein signaling modulator 2
GR	Glucocorticoid Receptor
GRIA1	glutamate ionotropic receptor AMPA type subunit 1
GRIN1	glutamate ionotropic receptor NMDA type subunit 1

HMCN1	hemicentin 1
IGSF5	immunoglobulin superfamily
KCNA2	potassium voltage-gated channel subfamily A member 2
KCNK9	potassium two pore domain channel subfamily K member 9
KCNQ4	potassium voltage-gated channel subfamily Q member 4
LRCH1	leucine rich repeats and calponin homology domain containing 1
LRP1B	LDL receptor related protein 1B
MCAT	malonyl-CoA-acyl carrier protein transacylase
METTL13	methyltransferase like 13
MFAP3	microfibril associated protein 3
NAV2	neuron navigator 2
NCOR1	nuclear receptor corepressor 1
NKAIN3	sodium/potassium transporting ATPase interacting 3
NLE1	notchless homolog 1
NMDA	N-methyl-D-aspartate
NRN1	neuritin 1
OPRM1	micro-opioid receptor 1
P2RX2	purinergic receptor P2X 2
PARP1	poly(ADP-ribose) polymerase 1
PAX3	paired box 3
PCNP	PEST proteolytic signal containing nuclear protein
PGRMC1	Progesterone Receptor Membrane Component 1
PLEKHG3	pleckstrin homology and RhoGEF domain containing G3
PMFBP1	polyamine modulated factor 1 binding protein 1
POLK	DNA polymerase kappa
POMC	pro-opiomelanocortin
POU1F1	POU domain class 1
POU3F3	POU domain class 3
PRL	prolactin
PRLR	prolactin receptor
PTK2	protein tyrosine kinase 2
PTPRC	protein tyrosine phospatase
PTPRZ1	protein tyrosine phosphatase
RBFOX1	RNA binding fox-1 homolog
RNF144A	ring finger protein 144A

RSAD2	radical S-adenosyl methionine domain containing 2
SEPT7	septin 7
STS	Steroid Sulfatase
TGFA	transforming growth factor alpha
TNN	tenascin N
TRPM3	transient receptor potential cation channel subfamily M member 3
TRPV2	transient receptor potential cation channel subfamily V member 2
TSYPL1	testis-specific Y-encoded-like protein 1
TSYPL4	testis-specific Y-encoded-like protein 4
TTC19	tetratricopeptide repeat domain 19
TTL	tubulin tyrosine ligase
TTLL11	tubulin tyrosine ligase like 11
TTR	Transthyretin
UBN1	ubinuclein 1
WDR47	WD repeat domain 47
XYLT1	xylosyltransferase 1
ZZEF1	zinc finger ZZ-type and EF-hand domain containing 1

Contents

List of Figures

List of Tables

1 Introduction

1.1 Maternal infanticide in *Sus scrofa*

In species that care for their young, early maternal behaviour is usually characterized by a phase of bonding with the offspring shortly after birth. In some cases the bonding will not occur and the response of the mother to its litter can result in carelessness or abandonment. For some species this can take a more extreme form, an aggressive response to the litter resulting in harm or death of the offspring. For commercial animal breeding, offspring survival is an important economical factor. It is reported in the US that 15% of piglets are lost before weaning, which resulted in an estimated loss of revenue ranging from 130 to 330 million dollars in the year 2000 [1]. Although the proportion of losses linked to infanticide is not well studied, the percentage of savaging dams generally varies between 5 and 15% [2, 3, 4, 5, 6], and death caused by infanticide according to some studies represent 11.2% of pre-weaning death [6]. Therefore, the infanticide cost to the industry could be between 14 to 36 millions dollars a year for the US alone.

Maternal infanticide (also called savaging) is a behaviour affecting sows shortly after giving birth, usually within twenty four hours after parturition (childbirth), resulting in the loss of one or more piglets. The sow attacks the piglets by biting them to death. Infanticide is not specific to domestic pigs and also affects wild boars raised in captivity [7].The etiology of this behaviour is complex and investigations to identify its causes have concluded that several potential factors can contribute to savaging. The main factors are environment, maternal experience and breed, the latter suggesting a genetic component. Studies have found that different breeds of domestic pigs have different incidences of aggressive behaviour [2]. Wild boar lines are also affected at different levels as one line was shown to be highly aggressive compared to others [7]. These differences in the levels of occurrence of infanticide in animals with a different genetic background support the possibility of a genetic component to this behaviour.

The contribution of environmental factors to the aetiology of aggressiveness has been the subject of a number of studies [6, 8, 9, 10, 11]. They show that the space and the richness of the environment available to the pig has an impact on the occurrence of infanticide behaviour, with barren environments tending toward more aggressive behaviour[12]. Larger pens and richer environments result in less aggressive behaviour, which could be the result of more exploratory behaviour that could help to alleviate stress. However some studies have shown that open pens have a higher incidence of infanticide [8]. This could be due to the sow having more access to the litter than in a farrowing crate and therefore more opportunity to attack its piglets.

Another environmental factor impacting infanticide is lighting of the pen. Constant lighting resulted in lower level of savaging [6]: however this contradicts earlier findings [3]. According to the former study, lighting could help the gilt (first time mothers) to be more aware of the piglets and therefore respond more positively to them.

Behaviour toward humans in contact with the sow has been investigated as a potential early sign for savaging behaviour [8]. This study looked at the impact of the behaviour of the sow towards its handler as a predictor of infanticide, investigating the theory that sows that are more aggressive towards humans might be more likely to kill their progeny. The conclusion was that aggressive behaviour toward the human handler has little correlation on the level of aggression directed toward the piglets. However "shy" behaviour of the dam toward humans can be a predictor of potential aggression. "Shy" gilts are more likely to be aggressive toward their litter than "bold" gilts. However this type of behaviour could also be the result of social hierarchy in pig groups, as "shy" guilts might be picking on weaker individuals (the piglets) as they are bullied by the more dominant sows. There is some suggestion of this behaviour in a study which used a resident intruder test [13], although not significant, there were observations that the smaller pigs were victim of aggression more often.

The posture and attitude of sows before parturition was also investigated as a potential early warning sign for a savaging event. Infanticide sows tend also to be more restless before and after parturition [5], which combined with the shy-bold trait suggests that infanticide is not triggered by parturition but predetermined. Considering these observations, a lack of experience with litter and birth could be a contributing factor as gilts are more likely to kill [14, 5]. It has also been confirmed that pigs can be conditioned in their behaviour by past experience [15, 11]. Although these factors contribute to offspring behaviour only for the few weeks after birth, they show that pigs can be conditioned to some extent by their experience in a given environment. Not all environmental factors have an impact on aggression, e.g. seasons and feeding have little impact [3].

Another potential cause of infanticide behaviour could be linked to endocrine levels, such as steroid levels in the sow before and just after birth. A study in gilts [9] concluded that there is no clear link between the infanticide behaviour and levels of hormones linked to birth, although the data demonstrated a higher (but not significant) level of oestradiol over progesterone a day before birth for gilts that had to be sedated because of their behaviour after parturition. This suggests that there might be an effect of the concentration ratio of hormones such as oestradiol and progesterone on infanticide.

The influence of inheritance in this behaviour has also been studied and it has been suggested that aggressive behaviour towards other animals is a heritable trait, with an estimated heritability for different breeds of between 0.4 and 0.9, with a strong genetic component. [16, 4, 3]. It is likely that the selection of meat traits for some markets has contributed to the selection of more aggressive traits resulting in a higher rate of infanticide in some breeds. In some Asian breeds there are almost no occurrences of infanticides. The desired meat traits for the Asian market is very different from the European ones. Furthermore, the selection of dedicated sires lines resulted in a selection for specific meat traits while dam lines are selected for litter and maternal trait. The statistics linked to maternal infanticide are reflective of this selection, in the study by Quilter et al [2], the two lines with a higher rate of infanticide are the two sire lines (line D and H) while the other two are dam lines (line B and C). In another study by Quilter et al [17], it was observed (unpublished) that the meat quantitative trait loci were close to the quantitative trait loci that were found in the linkage study. This observation suggest a strong genetic component to the infanticide trait.

Given that there is no good predictor of a sow's potential for infanticide other than some behaviours being more prevalent in savaging sows, such as restlessness and "shy" behaviour observed during an approach test. However there is evidence that the infanticide trait can be linked to heritability and varies between different pig breeds, therefore the genetic heritage and genomic composition has an influence on this trait. In order to get a better prediction of predisposition to savaging, the present study proposes to look more closely at the genome contribution to infanticide behaviour in the pig.

1.2 Method for the study of genetic markers.

1.2.1 Linkage analysis.

The study of genetic heritability started with Mendel's laws enacted from his work. These laws are the law of segregation, the law of independent assortment and the law of dominance. The law of segregation states that alleles will segregate during gamete formation and that each gamete will carry only one allele for each gene. The law of independent assortment states that alleles are transmitted independently from each other, and the law of dominance defines the nature of each allele in regards to the phenotype, either dominant or recessive.

The law of independent assortment was proven not to hold for every locus shortly after Mendel's work was completed. Further work on the heritability of traits by Morgan [18] and Punnet [19] challenged the rules discovered by Mendel, especially the law of independent assortment. This led to the discovery of genetic linkage, traits can be transferred together, as a single heritable unit, if they are in close proximity on the same chromosome. These discoveries gave rise to genomic methods to map the genome and eventually to the study of the genetic origins of diseases. One of the advantages of genomic methods is that they can be applied for a range of phenotypes, they do not require a particular type of disease to work other than having a genetic component and they can be applied to a wide variety of species. These methods are based on the fact that during meiosis recombination events happen, leading to cross over between homologous chromosomes. The frequency of recombination between loci is correlated, amongst other factors, to the distance between them; the greater the distance the more likely a recombination will happen.

Linkage equilibrium happens when loci are transmitted independently during gamete formation while Linkage Disequilibrium (LD) defines their joint transmission: the closer the loci the more likely they will be transmitted together. While a formal definition of Linkage Disequilibrium was not made until the 1960s [20], this principle lead to the discovery of genetic markers and was used to construct genetic maps, also called linkage maps. The first map of markers was constructed in the 1980s using Restriction Fragment Length Polymorphisms (RFLP) [21, 22] and later with microsatellites [23]. Once the presence of polymorphic markers was discovered on the human genome, analysing and typing those markers enabled the study of their transmission in order to identify the presence of a link between a given marker and a gene or genomic location.

The study of the transmission of markers to the next generation is carried out using linkage data. It usually consists of one or more families with marker data available from several generations (at least two). Several key parameters are needed to study the linkage of traits. These are the penetrance of the trait, dominance of the allele and its inheritance. These parameters are discussed in more detail below. Using this information and parameters, linkage based methods first extract the inheritance patterns from the pedigree and from this information determine causation between loci and the trait of interest.

1.2.1.1 Log Odds Ratio for single markers

The first methods developed to test for linkage are parametric methods and use the recombination fraction as the parameter for the testing procedure. The recombination fraction is defined as θ, and represents the probability of recombination between two loci. Two loci are in linkage equilibrium when $\theta = 0.5$ and are in linkage disequilibrium when $0 < \theta < 0.5$.

In 1955 Morton [24] defined a method to test for deviation from $\theta = 0.5$ called the sequential test for the detection of linkage. Given a set of observations of recombination $y_1, y_2, y_3....y_n$ and a recombination fraction θ for these loci, a random variable $f(y_{i,}\theta)$ for $i = 1, 2, 3, ...n$ is defined, which represents the probability of obtaining the current data set given θ. The null hypothesis H_0 is defined when $\theta = 0.5$, there is no linkage between the loci, the alternate hypothesis H_1 is defined as $\theta = \theta_1, 0 < \theta < 0.5$. Morton then defines his test parameter as:

$$z = \log \tfrac{f(y_i,\theta_1)}{f(y_i,0.5)} \tag{1.1}$$

The ratio z is called the Log Odds Ratio (LOD), where $f(y_i, \theta_1)$ represent the odds of obtaining the current pedigree given the recombination fraction θ_1, $f(y_i, 0.5)$ represent the odds of obtaining the current pedigree under H_0.

In simpler terms, it is the ratio of the probability of having the observed birth sequence if $\theta = \theta_1$over the probability of having the observed birth sequence if $\theta = 0.5$ (H_0). In practice the value θ_1 to test is unknown, Morton suggested trying several values for θ_1 (arbitrarily chosen) and the one scoring the highest (Maximum Likelihood) on z is the best approximation of θ_1.

In practice to compute the LOD for a single marker, representing a single typed genetic marker linked to a disease loci, the probability of the recombination event given the recombination fraction θ is defined as $p_r = (\theta/2)^r((1-\theta)/2)^{nr}$ and the probability under the null as $p_{nr} = (0.5/2)^N$ with r the number of recombinant individuals, nr the number of non recombinant individuals, N the total number of descendants in the generation studied ($N = r + nr$). The equation from Morton 1.1 can be written as a $log(p_r/p_{nr})$. Morton suggested that a test value should be above 3, corresponding to a likelihood of a 1000 to one that the marker is linked to the disease. It is considered as enough evidence for significant proof of linkage. This value is justifiable if the size and composition of the human genome is taken into account. To be linked, two loci need to be on the same chromosome and relatively close together, therefore this gives a prior odd of linkage between two random loci of around 1:50. Considering that a likelihood ratio of 1000:1 is sufficient evidence for linkage, given a prior of 1:50 this gives an odd of 20:1 in favour of the linkage being significant. To identify candidates a $log(p_r/p_m)$of 3 can be considered as significant and enough to reject H_0 and indicate linkage between the trait and the marker *[24]*.

An example of how the LOD is calculated is given in Table 1.1 for the pedigree in Figure 1.1. In this case we can consider that chromosomal phasing is known as the grand parents are genotyped. Chromosomal phasing or phase is defined as the transmission of alleles from the maternal or paternal chromosome. Phase is known when it is possible to determine which alleles came from the father or the mother. If the phase is unknown (grand parents not genotyped) the calculation must take into account all the possible phases from the parental genotypes. This results in a loss of power for the test as shown in Table 1.2. The example in Figure no. 1.1; is for a single family, in general a study will include several families. The method for LOD calculation from multiple pedigrees is simply to sum the individual LOD scores across all of the pedigrees. Therefore the more families present in a study, the more likely it will result in a high LOD score if linkage is present. Adding more families therefore increases the power of the test.

Recombinant fraction	p_r	p_{nr}	$log(p_r/p_{nr)}$
0.01	3.002E-04	9.766E-04	-0.512
0.05	1.273E-03	9.766E-04	0.115
0.1	2.050E-03	9.766E-04	0.322
0.2	2.560E-03	9.766E-04	0.419
0.25	2.472E-03	9.766E-04	0.403
0.3	2.251E-03	9.766E-04	0.363
0.4	1.620E-03	9.766E-04	0.220
0.5	9.766E-04	9.766E-04	0

Table 1.1: LOD calculations for the pedigree represented in Figure 1.1. p_r and p_{nr} are calculated with $r = 1$ and $nr = 4$. The maximum is reached for $\theta = 0.2$ which is close to the θ that can be estimated from the data. One recombination over 5 individuals give us an empiric recombination probability of 0.25. However the maximum LOD score is low for this simple example, so H$_0$would not be rejected in this case.

Phase 1(3:C, 2:B)				Phase 2 (3:B, 2:C)				
θ	p_r	p_{nr}	$log(p_r/p_{nr)}$	θ	p_r	p_{nr}	$log(p_r/p_{nr)}$	LOD combined score for both phase
0.01	3.002E-04	9.766E-04	-0.256	0.01	3.032E-06	9.766E-04	-1.254	-1.510
0.05	1.273E-03	9.766E-04	0.058	0.05	6.698E-05	9.766E-04	-0.582	-0.524
0.1	2.050E-03	9.766E-04	0.161	0.1	2.278E-04	9.766E-04	-0.316	-0.155
0.2	2.560E-03	9.766E-04	0.209	0.2	6.400E-04	9.766E-04	-0.092	0.118
0.25	2.472E-03	9.766E-04	0.202	0.25	8.240E-04	9.766E-04	-0.037	0.165
0.3	2.251E-03	9.766E-04	0.181	0.3	9.647E-04	9.766E-04	-0.003	0.179
0.4	1.620E-03	9.766E-04	0.110	0.4	1.080E-03	9.766E-04	0.022	0.132
0.5	9.766E-04	9.766E-04	0	0.5	9.766E-04	9.766E-04	0	0

Table 1.2: LOD calculation for the pedigree represented in Figure 1.1 when the grand parent data are not available, therefore phase is unknown. In this case the 2 possible phases from the disease parent need to be considered, e.g. 3:C/2:B or 3:B/2:C. Both have a probability of 1/2. In the first case the calculations are the same as shown in Table 1.1; the final LOD score is divided by 2 to take in account the probability of this phase. For the second phase, we have 2 recombinations in this case rather than one, the maximum LOD is again observed for to be matching the estimated θ(2/5=0.4). The combined LOD score is lower, as not having the phase further diminishes the power of the test.

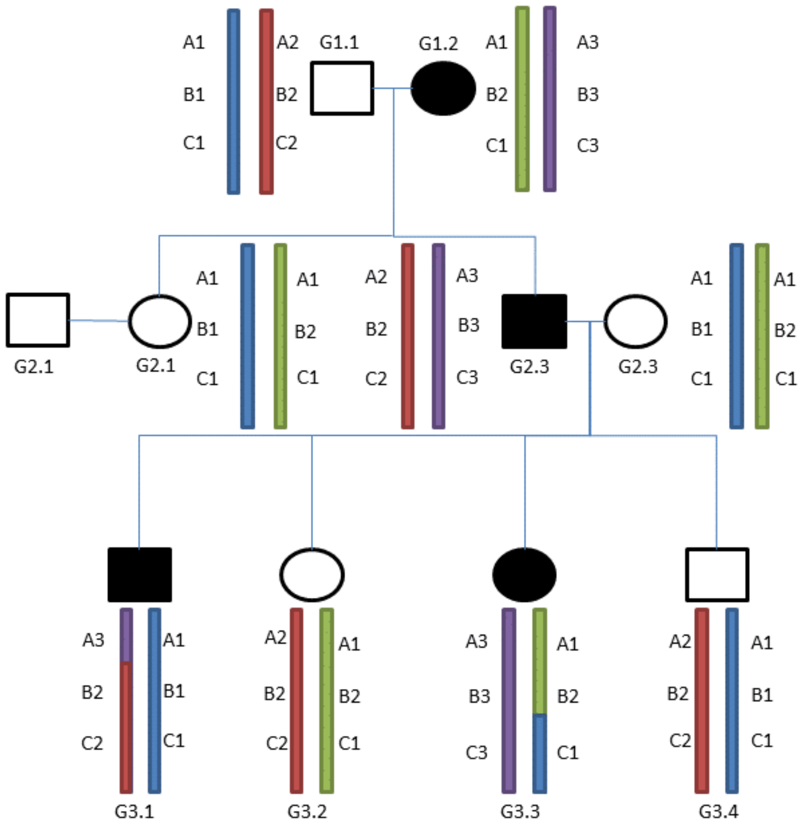


Figure 1.1: Linkage Example. Phase is known as the grand parents have been typed for all 3 markers. Each marker, A, B and C has three different alleles noted 1, 2 and 3. The generation a individual belongs to is noted G1,G2 or G3 and each different individual is denoted by a number after the generation. Total penetrance is assumed and the allele for the disease is dominant. In this simple model it is easy to work out that the disease is linked to allele A3, given the one recombinant individual on the third generation.

1.2.1.2 Log Odds Ratio for multiple markers

For multiple markers analysis, the LOD is calculated for each marker locus and graphs are constructed in order to report the LOD for each position of the marker map. Several algorithms have been developed in order to facilitate the calculation of the LOD for large pedigrees and multiple loci.[22, 25]. The algorithm developed by Lander and Green [22]allows the computation of LOD for a large number of markers given small pedigrees while the Elston algorithm [25] allows the computation of LOD for large pedigrees for a low number of markers by modelling the transmission of alleles down the generations using transmission matrices.

The Elston algorithm calculates the probability of an individual genotype conditional on the probability of its parents, starting from the most recent generation of the pedigree and repeating the process for each generation. The computational requirements increase linearly with the number of individuals, however the process becomes very costly as when the number of markers increases, the computational requirements increase exponentially. The Lander and Green algorithm uses a Hidden Markov Model in order to model the inheritance distribution by conditioning it on the observed genotypes, allowing computation of the LOD for small pedigrees but a large number of markers. With this approach the computational requirements are linear with the number of markers but augment exponentially with the number of individuals in the pedigree. GENEHUNTER [26] is a popular implementation of the Lander and Green algorithm, capable of multipoint linkage calculation with partial pedigree information.

The parametric methods of linkage analysis are however limited to good pedigrees and diseases with high or complete penetrance; incomplete penetrance and missing information have a high impact on the power of the sequential test. Penetrance is defined as the genetic value used to describe the impact of the genetic component on the phenotype of interest; a higher penetrance for an allele results in a higher risk of displaying the phenotype. If we define genetic penetrance of a disease causing allele A as γ ($\gamma > 1$); then for an additive model the risk of the disease is increased by a factor of γ fold with one A allele and by a factor 2γ fold for two A alleles. For the recessive model, the risk is increased by γ fold only for the AA genotype, but for a dominant disease the risk increases by γ fold in the presence of one or two A alleles [27].The mode of inheritance can also have an impact on tests using the recombination fraction. If the mode of transmission is not known or wrongly specified, it will impact on the power and the results of the test, although it might vary depending on the data [28].

1.2.1.3 Non parametric linkage test

To avoid the type of issue highlighted before, non parametric linkage tests were developed to evaluate linkage without estimating the recombination fraction [26, 29]. Although non parametric tests are less powerful than parametric tests, they are better suited when the inheritance mode of the disease is unknown. The most efficient methods for this type of analysis use alleles shared between individuals in the pedigree inherited from a common ancestor; such alleles are said to be in Identity By Descent (IBD). Other methods will use Identity By State (IBS), alleles that are shared between individuals in the pedigree. It is obvious that an allele in IBD is also in IBS but not vice versa. Methods using IBD are more powerful than methods using IBS as more information can be captured in order to determine linkage. One of the first non parametric approaches was the Affected Pedigree Member (APM)[30], using IBS between pairs of affected members of the pedigree. To compute the test statistics, a score is defined by counting the alleles shared in IBS between a pair of individuals. The score is calculated by giving a 1 if an allele is identical between the pair (so is IBS at least) and 0 if not. The following formula is defined for a pair of two affected individuals labeled i and j and with x and y representing respectively the maternal and paternal alleles for each individual:

$$Z_{ij} = 1/4\delta(G_{ix}, G_{jx})f(p_{G_{ix}}) + 1/4\delta(G_{ix}, G_{jy})f(p_{G_{ix}}) \\ + 1/4\delta(G_{iy}, G_{jx})f(p_{G_{iy}}) + 1/4\delta(G_{iy}, G_{jy})f(p_{G_{iy}}) \quad (1.2)$$

With $\delta(G_1, G_2)$ defining the Kronecker delta such as $\delta(G_1, G_2) = 1$ if G_1 is in IBS with G_2 and 0 otherwise. $f(p_G)$ represents a function of the allele frequency of the allele studied in the population, usually in the form $1/p$ or $1/\sqrt{p}$, thus allowing us to weight the score of an allele according to its frequency, p, in the population. For a given pedigree the score for an allele using the APM will be the sum of all the Z_{ij}. This method is usually used to combine scores from several families in order to generate a global statistic for an allele using several pedigrees, if Z_m is Z for the m^{th} pedigree, then we have the test statistic:

$$T = \sum_m \frac{w_m[Z_m - E(Z_m)]}{\sqrt{w_m Var(Z_m)}} \quad (1.3)$$

Where w_m is a positive weight, usually derived from the number of affected individuals and $E(Z_m)$ the expected value of Z_m. Given the central limit theorem, T will follow a normal distribution and it is therefore possible to derive a p-value from it.

This method is fairly restrictive as it only uses pairs of individuals and uses IBS information instead of the more powerful IBD information that may be inferred from pedigrees. It also ignores data that could be used to determine the status of the alleles, IBD or IBS [26]. Other methods of non parametric testing have been developed, more appropriately for looking at large pedigrees or combining information from several pedigrees [26, 29]. They use a scoring method by looking at allele sharing by IBD or at all the permutations possible that could lead to allele sharing by IBD, usually by sampling alleles and counting the number of times a founder allele is shared with subsequent generations. For example Whittemore and Halpern [29] define two different metrics for scoring a configuration of alleles in IBD. First they define a IBD configuration, denoted ϕ, representing the configuration of alleles in IBD between individuals. If $G = (G_{1x}, G_{1y}, ..., G_{nx}, G_{ny})$ are $2n$ integers representing the inheritance vector for all the alleles, where x and y define the maternal and paternal allele respectively, as seen before. The sharing of IBD can then be assessed between siblings using the following method: given a locus with alleles a, b, c, the allele configuration between two individuals can be noted as (a, b, a, b), representing the sharing of both paternal and maternal alleles, whereas (c, b, a, b) would equate sharing only the paternal allele. If two sequences G and G' only differ by the order of their paternal and maternal alleles, as in the example(a, b, a, b) and (a, b, b, a) we define the IDB configuration as $\phi = [abab]$ with (a, b, a, b) and (a, b, b, a) as representative of ϕ. Generalizing ϕ we have $\phi = [G_{1x}G_{1y}...G_{nx}G_{ny}]$ representing all the possible IBD configurations between the $2n$ alleles of a pair of individuals. The scoring functions for this ϕ, τ_p is defined as

$$\tau_p = S_p(\phi) = \frac{2}{n(n-1)} \sum_{1<i\leq j<n} f_{ij}(\phi)$$

$$f_{ij}(\phi) = 1/4[\delta(G_{ix}, G_{jx}) + \delta(G_{ix}, G_{jy}) + \delta(G_{ix}, G_{jy}) + \delta(G_{ix}, G_{jy})] \quad (1.4)$$

for a pair of individuals, which is very similar to the definition of (1.2) without taking into account the frequency of the alleles.

In order to use more than pairs of individuals to score IBD in the pedigree, define $u = (u_1, u_2, ..., u_n)$ where u_i is either G_{ix} or G_{iy}, for each ϕ there are 2^n vectors u. Let $h(u)$ be the number of non trivial permutations of the u_i in u that leave it unchanged. The more IBD is shared between individuals the larger $h(u)$ will become. The scoring function, τ is defined from $h(u)$ as its average:

$$\tau = S_{all}(\phi) = \frac{1}{2^n} \sum_u h(u) \quad (1.5)$$

For two individuals the results of the two scores will be identical. For more than 2 individuals ($n > 2$) the two functions will differ, τ is more powerful when the disease is dominant while τ_p is more powerful when the disease is recessive as the power calculation in [29] have shown. Those two scoring methods have more power than the APM test as they use IBD while APM relies on IBS.

The two scoring functions from Whittemore and Halpern [29] were refined by Kruglyak et al. [26], introducing a normalised score for both methods $Z = [S(\phi) - \mu]/\sigma$ where $S(\phi)$ can be either $S_p(\phi)$ or $S_{all}(\phi)$ and μ and σ are respectively the mean and standard deviation of the scoring function under a uniform distribution $P_{uniform}$, representing the uniform distribution over all the possible inheritance configurations of the inheritance vector. The mean and standard deviation can be calculated by enumerating all the possible inheritance vectors. Under H_0, Z has a a mean of 0 and a standard deviation of 1. To combine data from several pedigrees Kuglyak et al. defines Z for multiple pedigrees as $Z = \sum_i^m \delta_i Z_i$ with δ_i a weighting factor such as $\sum \delta_i^2 = 1$, so that under the null hypothesis, the mean and standard deviation for Z are still 0 and 1. Taking $\delta_i = 1/\sqrt{m}$ is a good compromise as it seems to perform well for both small and large pedigrees. This method has been implemented in the software GENEHUNTER[26].

While linkage had a lot of success initially, it was employed as a method to identify genes and markers linked to a simple disease from a genomic perspective, such as dominant or recessive traits linked to a few locations and with good penetrance. It had good success identifying the susceptibility of developing breast cancer linked to the *BRCA* family [31, 32], the link between lipoprotein-cholesterol and chromosome 9 and was also used to identify loci linked to early onset Alzheimer's disease [33, 34] . The successful linkage studies focused on those diseases with a strong correlation between the genetic component and the phenotypic data. Linkage proved to be a very powerful approach to investigate

Mendelian disease with a single genetic component. However the method is a lot less powerful when investigating a more complex disease [35] and for several reasons [36]. This is particularly true for diseases caused by high frequency, low effect size polymorphisms. The fact that most methods rely on allele sharing is a problem when the penetrance is limited, since an individual might carry the allele but if the risk associated is low, its contribution to the phenotype may not be detected. This individual will be classified as unaffected while still carrying risk alleles, and will be removed from the analysis. The removal of these individuals will results in a loss of power for the test. The high frequency of such alleles is also an issue as it will make it more difficult to effectively follow its inheritance pattern if it is common in the population; these high frequency allele can enter the pedigree via multiple founders. To detect common markers that are linked to the disease, it is possible to mitigate the impact of the latter problem by typing more families and increasing the size of the study but this is not always practical.

Linkage test are powerful to study high penetrance and low frequency disease, also called Mendelian disease. However in the case of more common diseases, caused by high frequency low penetrance alleles, the inheritance pattern of the alleles is harder to link back to the disease as the individual makers contribute less to the phenotype. A group of several markers is more likely to significantly contribute to the phenotype of interest. Therefore, to study the more common genetic diseases that are more frequent in a population, a new approach was needed.

1.2.2 Genome Wide Association Studies (GWAS)

While linkage studies identify loci using heritability and recombinations within a pedigree, an alternative is to study linkage disequilibrium at the population level, using unrelated individuals to identify loci associated with a particular phenotype. LD is stronger for loci close together in the genome but is also influenced by the nucleotide composition and position of the markers on the chromosome; therefore it is heterogeneous along the genome. By mapping LD and selecting markers it is possible to study the association between markers (and therefore of the region/genes they represent) and a particular phenotype at the population level. This method is called Genome Wide Association Study or GWAS. Linkage studies started to show limitations in their power to identify markers when a disease had reduced penetrance, resulting in difficulty for the recruitment of a sufficient number of samples to achieve good power [37]. This is especially true for complex and common diseases where several loci in the genome can contribute to the genotype of the disease. By comparing allele frequencies between control and case (affected) populations it is possible to test for loci associated with the phenotype of interest. However this approach can have some drawbacks depending on the disease studied. For example A disease such aS Beta-thalassemia can be challenging to study using GWAS due to genetic heterogeneity *[38]*. The phenotype of this disease can be the result of more than 200 disease causing mutations. Individuals classified as affected might have different genetic loci involved which are all resulting in the same phenotype. Furthermore, the different genetic causal loci might be mixed within a population or segregating in different populations. If different loci are causing the phenotype within the same population, it will considerably reduce the power of a GWAS on this population. If this happens between different populations, different loci might be identified in each population despite them sharing a common phenotype .

1.2.2.1 Transmission Disequilibrium Test

The first method to study association is still very close to linkage and is the Transmission Disequilibrium Test or TDT proposed by Spielman et al. [37]. It tests for linkage in the presence of association in parent-affected child trios by looking at which combinations of alleles are transmitted to the affected child and which are not. Previous approaches were focusing on the transmission of individual alleles instead of the combination of alleles. Suppose that we have a disease allele D_1 and normal allele D_2 and a marker locus with co-dominant alleles M_1 and M_2. Consider the Table 1.3 of transmitted and non transmitted alleles. It was shown in [39] that the probability associated with each value in Table 1.3 is as shown as in Table 1.4 with m the frequency of the M_1 allele, p the frequency of D_1, θ the recombination fraction and δ the coefficient of LD. As the TDT is testing for linkage ($\theta = 1/2$) the only two terms of interest from Table 1.3 are b and c as they are the only ones for which the probability function contains θ. For this test only heterozygous parents typed in the pedigree are informative. If H_0 is true ($\theta = 1/2$, no recombination) then it is trivial that $E(b) = E(c)$ using the expression of probability in Table 1.4 and this is independent of the values of m, δ and p. We can then define a χ^2 to reject H_0. Any χ^2 test to be valid is in the form of $\chi^2 = (u-v)^2/Var(u-v)$. As in most statistical tests, the true variance is impossible to calculate, therefore the test for b and c can be written $\chi^2 = (b-c)^2/estimate(Var(b-c))$. If $\theta = 1/2$ then the contributions of the heterozygous parent are independent. In this case, the test is the 'McNemars test' and the statistic for the TDT is

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

	Non Transmitted allele		
Transmitted allele	M_1	M_2	Total
M_1	a	b	$a+b$
M_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

Table 1.3: Combination of Transmitted and Non Transmitted Marker Allele M_1 and M_2 among $2n$ Parents of n Affected Children [37]. a, b, c and d represent the number of parents in each categories, e.g. b is the number of heterozygous parents M_1M_2 that transmitted allele $M_1$and not allele M_2 .

	Non Transmitted allele		
Transmitted allele	M_1	M_2	Total
M_1	$m^2 + (m\delta/p)$	$m(1-m) + [(1-\theta-m)\delta/p]$	$m + [(1-\theta)\delta/p$
M_2	$m(1-m) + [(\theta-m)\delta/p]$	$(1-m)^2 - [(1-m)\delta/p]$	$1-m-[(1-\theta)\delta/p]$
Total	$m + (\theta\delta/p)$	$1-m-(\theta\delta/p)$	1

Table 1.4: Probability of Combinations of Transmitted and Non Transmitted Marker Allele M_1 and M_2among $2n$ Parents of Affected Children

Another alternative to the definition of χ^2 is to consider the test as the following. The parents transmission of their alleles can be described as a series of $2n$ multinomial trials with four possible outcomes as described in [40]. The outcomes have probability x, y, y and z with $x+2y+z=1$, which correspond to our case of $\theta = 1/2$ as $E(b) = E(c) = 2ny$. If a, b, c and d are the number of possible outcomes in the four categories then we have $Var(b) = Var(c) = 2ny(1-y)$ and $Cov(b,c) = -2ny^2$. Then $Var(b-c) = Var(b) + Var(c) - 2Cov(b,c) = 4ny$. Therefore the best estimate of the variance is $b+c$ so the TDT statistic is $\chi^2 = \frac{(b-c)^2}{(b+c)}$. The TDT procedure described here will work on a parent-affected child trio but can be extended easily to bigger families [37]. The TDT is more sensitive than allele sharing methods traditionally used for linkage and can make good use of family trios and families with multiple affected siblings. The only drawback to the TDT is the need for association between the marker at the population level in order to be able to assert linkage, it is however resilient to population stratification and can be used as an (secondary) independent test to confirm GWAS results.

1.2.2.2 Parent of Origin test: application of the TDT test.

The TDT to test can be used for preferential transmission of alleles from the father or the mother of the affected offspring. The test procedure is the same as the standard TDT but only the heterozygous mother's or father's alleles are considered in order to test for preferential transmission of some markers. In this study the test was used because it had been shown that a dam to daughter effect is present [3]. This test was used to identify any preferential transmission from the dam to her daughter that might be linked to infanticide.

1.2.2.3 Population admixture

Real associations studies, performed to investigate association between loci and a phenotype, are performed on unrelated individuals and not on family data and therefore do not look for the presence of linkage with the disease allele. The tests are performed between two populations of unrelated individuals, one of control and one of affected individuals also referred to as 'cases'. The selection of both populations is important as the results of the analysis can be biased by spurious associations if there is a sub-structure in the population such as common ancestry or a particular ethnic group forming part of one of the populations [27]. This is called populations stratification, often the result of groups of different genetic origin forming the population. Another type of bias can be caused by admixture in the population, usually due to genetic mixing with populations of different genetic origin in recent history, introducing new alleles and shifting the allele frequencies of the original population. Population stratification and admixture can be assessed using Principal Component analysis (PCA) which decomposes the variance by maximizing it and minimizing the error using Eigenvectors. Sub populations are usually seen as clear outliers when plotting the largest components of a PCA.

Another important test to consider when looking at association is the Hardy Weinberg Equilibrium (HWE). If the frequency of the allele a in the population is p and the frequency of the disease allele A is q then the frequency of the genotypes in the population should be: p^2 for aa q^2 for AA and $2pq$ for the aA genotype. Any significant deviations from those proportions is indicative of admixture and/or stratification in the population. Testing for HWE can be done using a χ^2 test on a contingency table. Testing for HWE is performed on the control population generally as the case population can have markers in disequilibrium due to the genetic component of the disease [41]. In general the study of association focuses on common variants (markers) with a Minor Allele Frequency (MAF) in the population greater than 5% or 10% because of the sample size and disease penetrance needed to achieve power at lower allele frequency [42]. If an allele is associated with a causal loci, but its frequency in the population is low, it will not be well represented in the case population. It is unlikely that the association test will successfully identify this allele as associated with the disease, its contribution at the population level being too low. In order to use rarer alleles in association studies, more individuals need to be typed in order to have enough observations of this allele to find the association with the disease.

1.2.2.4 Association test

Testing for association can be done with a simple test on a contingency table of the genotype count. This can be done with a χ^2 test, Fisher Exact test [43] or Cochran-Armitage trend test [44, 45] to test for statistical significance for each marker tested. The Fisher exact test is more appropriate for data with a low sample counts while the Cochran-Armitage test is better suited to handle data when the dominance of the disease is not known [27]. The contingency table can be based on allele counts, Table 1.5, or on genotype counts, Table 1.6. The Fisher exact test on the allele counts table will be written as

$$\chi^2 = \frac{[\frac{(x_{11}+x_{12})!}{x_{11}!x_{12}!}] * [\frac{(x_{21}+x_{22})!}{x_{21}!x_{22}!}]}{(2n!/(x_{11}+x_{21})!(2n-x_{11}-x_{21})!} = \frac{(x_{11}+x_{12})!(x_{21}+x_{22})!(x_{12}+x_{22})!}{(2n!x_{11}!x_{21}!x_{12}!x_{22}!)}$$

the formula can be extended to test on the genotype count table.

The Cochran-Armitage trend test of association between disease and marker is calculated on the genotype counts table in the form of:

$$T^2 = \frac{[\sum_{i=1}^{3} w_i(Sn_{1i} - Rn_{2i})]^2}{[RS/n] * [\sum_{i=1}^{3} w_i^2 N_i(n - N_1) - 2\sum_{1=1}^{2}\sum_{j=i+1}^{3} w_i w_j N_i N_j]}$$

where the denominator is the decomposed variance of the numerator. This trend test can account for the various possible models of penetrance for the disease using the term w_i as a weight for the impact of the genotype on the phenotype. Following the penetrance model (if known) it is easy to choose the w factor for the trend test: for a recessive model the w can be set as $w = (0, 0, 1)$, for a dominant model it can be set as $w = (0, 1, 1)$ and for an additive model $w = (0, 1, 2)$.

Allele	a	A	Total
Case	x_{11}	x_{12}	$x_{11} + x_{12}$
Control	x_{21}	x_{22}	$x_{21} + x_{22}$
Total	$x_{11} + x_{21}$	$x_{12} + x_{22}$	$2n$

Table 1.5: Allele count table. Two alleles are present: 'A' and 'a', the alleles are counted for each population, case and control. x_{11}represents the total allele count of allele 'a' in the case population, $x_{21}$the count of that allele in the control population.$x_{12}$and x_{22} are the same but for the 'A' allele.

Genotypes	a/a	a/A	A/A	total
Case	n_{11}	n_{12}	n_{13}	$n_{11} + n_{12} + n_{13} = S$
Controls	n_{21}	n_{22}	n_{23}	$n_{21} + n_{22} + n_{23} = R$
Total	$n_{11} + n_{21} = N_1$	$n_{12} + n_{22} = N_2$	$n_{13} + n_{23} = N_3$	n

Table 1.6: Genotypes count table, the genotypes are counted for each population, given three values for each population for a given loci, $n_{11}$is the number of a/a genotypes in the case population, $n_{21}$the same in the control population, $n_{12}$is the number of a/A genotypes in the case population and $n_{22}$for the controls, $n_{13}$is the number of A/A genotypes in the case population and n_{23}in the control population. Summing each line we obtain S the number of genotypes for the case population and R for the control population. N_1, $N_2$and N_3 represent the number of genotypes for a/a, a/A and A/A respectively for both populations.

1.2.2.5 Multiple testing correction and significance level

Usually the tests are carried out on thousands of markers at the same time, which results in a high probability for false positive or type I error. This is because each test is supposed to be independent but markers can affect each other so this assumption is violated. To avoid or reduce type I error, multiple testing corrections are usually applied. The simplest method is the Bonferroni correction which takes the chosen threshold p-values and divide them by the number n of tests carried out (the number of markers used in the study). This method however is very stringent and often too conservative, so an alternative was proposed by Benjamini and Hochberg [46]. The p-values are ranked from the lowest to the highest, multiplied by the number of tests carried out and then divided by their rank. The last step is to make sure that the rank of the p-value is conserved. This is a more moderate correction as the rank of the p-value is used as a weight. It is widely used; however for GWAS analysis, permutations based methods are more popular and give better results [27]. Bootstrapping methods are the most popular and work by randomly distributing the label of “control” and “case” across the samples in the study. The association test is then performed and the p-value obtained. The process is repeated a number of times (usually a few hundred to a few thousand times) and a distribution of the p-values can be built against which the empirical p-value from the true population can be compared to assess for true significance. These approaches have the drawback of only working for the subset of markers studied and are only truly valid if a set of genome wide markers is available. The choice of the level of significance in human studies has been subject to some debate [47], the first consensus is to use 5×10^{-8} which is significant if all the SNPs discovered by HapMap [48] (~10^{6}) are tested and the level of significance is set at 0.05. The level of significance is subject to the population tested, the strength of association and the marker used for the analysis. Other work [47] suggests that SNP at 10^{-7} are worth investigating and that the 5×10^{-8}might be too stringent. While a lot of work has been done to work out the level of genome wide significance in humans, no guidelines are available for livestock; however one can assume that the level of significance should be similar as the genomes are fairly similar in size and number of markers. Most studies in the human use a large number of markers or markers that have been identified to specifically type a particular area of interest. A higher density of markers will help better type the genome as it will allow the analysis of smaller LD blocks, a more sparse spacing of the markers might results in smaller LD blocks not being typed adequately. This is an issue for this study as the array used as it does not have a large number of markers, see section about the microarray technology 1.5 and discussion about the array 3.1.3.

1.2.2.6 Genotypic Risk Ratio and Odds Ratio

Once association is established, it is interesting to look at the strength of the association; this can be done by two means. The Genotypic Risk Ratio (GRR) or Odds Ratio (OR). If we take $\pi_{AA}, \pi_{Aa}, \pi_{aa}$ as the disease penetrance for the AA, Aa, aa genotypes respectively, then the relative risk in relation to the AA genotype would be $\theta_{AA} = \pi_{AA}/\pi_{aa}$ and the relative risk for the Aa genotype would be $\theta_{Aa} = \pi_{Aa}/\pi_{aa}$. The GRR is rarely used as it requires a population to be selected prior to developing the disease in order to provide an estimate of the penetrance in the general population. Using a case-control study will negate the possibility of assessing the penetrance as the ratio of cases to controls is controlled by the investigator. A more common way of assessing the strength of association is to look at the allelelic or genotypic odds ratio. For the allelic odds ratio, the ratio is $OR_{Allele} = \frac{[x_{12}/x_{11}]}{[x_{22}/x_{21}]} = \frac{x_{12}x_{21}}{x_{22}x_{11}}$ using the value from Table 1.5. This ratio represent the odds of being affected when carrying the allele linked to the phenotype of interest. For the odds ratio in relation to the genotype it is important to separate the homozygote from the heterozygote genotype. Using the notation in Table 1.6 $OR_{AA} = \frac{[n_{13}/n_{11}]}{[n_{23}/n_{21}]} = \frac{n_{13}n_{21}}{n_{11}n_{23}}$ while $OR_{aA} = \frac{[n_{12}/n_{11}]}{[n_{22}/n_{21}]} = \frac{n_{12}n_{21}}{n_{11}n_{22}}$. The first ratio OR_{AA} represents the odds of being affected when carrying the AA genotype compare to the aa genotype. The second ratio OR_{aA} represents the odds for a heterozygote carrier.

1.3 Family based association test

The methods described above for association and linkage test are usually used when a complete data set is available, unrelated individuals for the association test and pedigree data for the linkage test. The TDT test relies on information from heterozygotes parents and cannot be used when one of the parents is missing. However it is often the case that not all the individuals in a pedigree can be typed for various reasons, e.g. lack of consent, individual is deceased (late onset disease) or was not typed during the study. In these cases in order to maximize the use of the data it is necessary to find information outside the traditional parent-infant trio.

Several methods have been proposed to solve this issue by using siblings in order to test for association in the presence of linkage [49], which compares the allele frequencies between affected and unaffected siblings. This test uses siblings in order to complete the data set if the parents are not available for typing. If linkage is not present then the allele frequency will be similar between non affected and affected siblings, only in the presence of linkage will it be possible to detect association between the marker and disease. By using a permutation procedure such as the Monte-Carlo permutation within families it is possible to test for a significant difference in allele frequencies between the two categories. A permutation approach is needed as the assumption of independence of sampling is not respected when siblings are tested. While this method is useful if siblings are available, it is not always possible to have data for the non affected sibling.

Another method makes it possible to infer the missing parental data from the affected offspring data or from any available parent data. The approach proposed by Rabinowitz and Laird [50] can test for association in the presence of linkage without the need to be adjusted for population admixture or the genetic model and can be used for pedigrees with missing marker allele information. The method uses a statistical principle called “conditioning on sufficient statistic for the null hypothesis”. It revolves around defining models that share the same conditional distributions given the statistics. This means that the computation for the p-value for all the models given the minimum statistics are the same. For this approach it is only required to compute the conditional distribution for the data given the minimal sufficient statistics for the null hypothesis, which will reduce considerably the computational burden. In the case of a nuclear family where the parents plus affected child trio are typed and their traits are known, the minimum statistics are defined as the trait and the parental genotypes. Under Hardy Weinberg equilibrium it is easy to calculate the conditional distribution for the markers. However the data might not always be complete, one or both parents could be missing. To be able to calculate the conditional distribution for the marker it is necessary for the trait status to be known for some individuals in the pedigree and in a similar fashion that some marker data are available for some members of the pedigree. It is not required that trait status and marker data are known for the same individual. In order to compute the conditional probability to test for linkage in pedigrees the following steps are proposed in [50]:

1. Find all the possible patterns of founder alleles that are compatible with the allele in the typed marker, this involves inferring missing genotypes in founders if needed.

2. Construct the pattern of typed marker alleles compatible with the set of founder alleles defined in step 1. Find the subset of patterns that are compatible with the observed typed marker in the available typed markers

3. For every pattern found in the first step and in the second steps, compute the ratio between:

    a) the conditional probability of the typed marker allele given the pattern for the founders marker allele, so in general the conditional probability for any potential pattern in the pedigree given the possible founder pattern, not restricted by the observed data.

    b) the conditional probability of the **observed** marker alleles given the pattern for the founder marker alleles.

4. For some subset compatible with the observed data found at step2, the ratio found at step 3 will all be the same for all the possible founders marker allele found in step 1, this is the set of outcome with positive conditional probability.

5. The conditional distribution is found by arbitrarily choosing any of the compatible pattern of founder makers found in step1 and computing the conditional probability for the marker allele typed given the founders pattern chosen and given that set is part of the positive conditional probability set described in step 4.

It is possible using this method to test for linkage following these steps. In order to test for association in the presence of linkage the procedure needs to be slightly altered. The presence of linkage implies that identity by descent will not be independent from the pattern of traits present in a pedigree because of the presence of linkage. Given that association is not assumed to be present, if founder genotypes are observed and IBD is available, the minimum statistic in this case is conditioned by the founder genotype and the IBD relationship within the pedigree. The patterns of founders described in the steps for the linkage test must now also take into account IBD, in a similar fashion for step 3, and the conditional probability needs to take into account the compatible marker and compatible IBD relationship. Any step referring to the pattern of founder alleles needs to incorporate relationship by IBD. Various examples of what the conditional distributions would be for linkage and association are given in [50].

The strength of this method is that it allows a test for association in the presence of linkage in a large number of families and for a large number of markers, therefore making it possible to use a data set with related information to find association.

1.4 Previous studies on pig infanticide

The link between the genome and the savaging trait has already been the subject of several studies using different approaches. These works include using a linkage approach, to identify regions of interest. The influence of gene expression was also investigated, comparing gene expression level in savaging sows against non savaging. Finally an association study comparing populations of infanticide sows against non infanticide was also carried out. In this part these studies will be summarised to give an overview of the current knowledge of the link between the genome and the maternal infanticide phenotype.

1.4.1 Pig as a model for puerperal psychosis

One interesting hypothesis investigated in one of these studies by Quilter et al [17] is that pig maternal infanticide could be a good model for human puerperal psychosis. This psychological disorder is an extreme form of baby blues that affects 1 in 1000 births. The symptoms include depression, suicidal thoughts, loss of appetite and poor bonding with the baby. It is usually linked to feelings of guilt, self worthlessness or hopelessness. Some extreme cases have unfortunately resulted in infanticide. Mothers with a history of bipolar disorder are more likely to develop puerperal psychosis after birth and it has been linked to other psychiatric disorders. Due to the similarity in the phenotype of the disease with maternal infanticide, genes implicated in maternal infanticide in the pig could be interesting targets to investigate the origin of puerperal psychosis in humans.

1.4.2 Linkage study

Several studies have already been done to assess the genetic origin of maternal infanticide in the pig. The first study by Quilter et al [17] used a microsatellite study to identify potential Quantitative Trait Loci (QTL) that might be linked to maternal infanticide. QTL are part of the genome which are associated with a phenotype, usually containing a gene. They defined large regions of the genome that can be further investigated. This particular study used 80 microsatellite markers and used a non parametric linkage analysis [51] on a hundred and nineteen animals from 11 different lines. Another linkage study was done on the pig by Chen et al [14], using a similar approach to investigate maternal infanticide. The animals for this study came from a White Duroc intercross with a Erhualian breed. The study investigated F2 sows across three successive rounds of farrowing, totalling 288 sows. The Erhualian breed is an excellent breeder with an average of 16 piglets per litter and has very good maternal behaviour. The percentage of reported infanticide dropped from litter to litter, with 12.8% for the first litter, 7,5% for the second and 4.5% for the third. Similarly to Quilter et al, a non parametric linkage analysis was used.

Both studies identified QTLs in common on chromosome X and chromosome 2. Quilter et al found an additional QTL on chromosome 10 while Chen et al found additional QTLs on chromosome 6, 14 and 15. Because of the lack of annotation for the pig genome at the time of these studies, both used synteny with the human genome in order to identify potential genes of interest.

Several QTL were found by both studies on chromosome X, SSCX, both on the short arm and the long arm of the chromosome. On the short arm, both studies identified a QTL located in a pseudo-autosomal region which includes the gene Steroid Sulfatase (*STS*) that codes for an enzyme processing sulfated steroids hormone precursors to oestrogen during the pregnancy. In human a deficit in *STS* is associated with an elevated risk attention deficit hyperactivity disorder, autism and social communication deficits [52]. In the mouse, aggressive behaviour has been linked to this region [53, 54]. *STS* concentration in the mouse liver was also linked to aggressive behaviour and was shown to modulate the behaviour [55, 54]. Chen et al proposed another candidate on the short arm, Glycine Receptor Alpha2 subunit (*GLRA2*) . *GLRA2* is an inhibitory neurotransmitter in the central nervous system. It is regulated by a specific glycine receptor which has been linked to the pathogenesis of schizophrenia and other psychiatric diseases [56].

On the long arm both studies identified a locus in the Xq2.2 region, Progesterone Receptor Membrane Component 1 (*PGRMC1*), which can bind several steroids including progesterone [57]. Progesterone blockade during late pregnancy leads to abhorrent maternal behaviour including infanticide [58]. Furthermore, increases in levels of progesterone and oestrogen have been linked to increased levels of aggression in the pig [9]. Another gene located in the syntenic region in human is *5HTR2C*, a stimulating phospholipase C and that has been linked to several abnormal behaviours. Serotonin regulates the release of dopamine via this receptor. Decreases in its level have been shown to alleviate depression by release of dopamine [59].This receptor has been linked to suicide [60], alcoholism [61], anorexia [62, 63] and the behavioural aberrations observed in Prader Willi patients (a genetic condition causing behaviour disorder and mental retardation) [64]. Others members of its family are associated with mood disorder and schizophrenia [65], making it a very interesting target.

For chromosome 2 of the pig, both studies identified QTLs at different locations, highlighting different genes. The study by Quilter et al identified a gene called *COX7C* (cytochrome subunit VIIc), which is located on the human syntenic region 15q14. It could be an interesting candidate as cytochrome is the terminal component of the mitochondrial respiratory chain and mitochondrial dysfunction has been associated with bipolar disorder and schizophrenia [66, 67]. This gene is also regulated by *YY1* [68], which has been shown to be a target for stress regulated pathway in neural degeneration [69]. The QTL identified by Chen et al is syntenic to the region of the Glucocorticoid Receptor (*GR*), located on chromosome 5q31.3 in human. This gene mediates the effect of glucocorticoid release in response to stress and the regulation of the hypothalamic-pituitary adrenocortical system using a negative feeedback mechanism. *GR* expression is deregulated in schizophrenia, depression and bipolar disorder [70, 71]. The syntenic region in human on chromosome 4 has also been linked to panic disorder and agoraphobia [72].

The QTLs on the other chromosomes also identify interesting genes, the QTL found on chromosome 6 by Chen et al has Albumin D-site-binding protein (*DPB*) as a candidate gene. Decreased expression of this gene in the mouse leads to increase susceptibility to generalized spontaneous and audiogenic epilepsies [73]. The QTL on chromosome 10 identified by Quilter et al has *PTPRC* (protein tyrosine phospatase, receptor type, C), also known has *CD45* has a candidate gene. It is a candidate for neuroinflamation which may exacerbate neurodegeneration and is found in conditions such as Alzheimer disease (AD) and Down's syndrome [74, 75]. For the QTL on chromosome 14 identified by Chen et al, the candidate gene in the corresponding human region is Alpah-2A adrenergic receptor (*ADRA2A*). It plays a criticial role in the regulation of neurotransmitter release from sympathetic nerves and adrenergic neurones in the central nervous system and has been linked to schizophrenia [76, 77]. For SSC15 a potential candidate close to the peak of linkage is the gene for the sodium channel voltage gated type II alpha which is involved in pain perception [78]. Another potential candidate for this region is glycoprotein M6A (*GPM6A*) which might be involved in stress induced hyppocampal alterations in psychiatric disorders such as schizophrenia [79].

1.4.3 Gene expression study

A follow up study on the linkage study done by Quilter et al [17] was done by Quilter et al [80] looking at the gene expression in the pig hypothalamus. The hypothalamus is an important part of the brain where parturition and maternal behaviours are regulated by neuroendocrine systems that originate from, and are coordinated by, the hypothalamic nuclei. The areas involved in this are the medial preoptic area, paraventricular nucleus and the supraoptic nucleus. These brain areas are responsive to stimuli such as sex steroids, oxytocin, prostaglandin F2α and prolactin (*PRL*) which have been characterised in pig and implicated in maternal behaviour. The study compared samples from infanticide individuals and samples from non aggressive individuals to test for differentially expressed genes, using two methods to test for over and under expression in the infanticide samples.

Several interesting candidates genes were identified as significantly regulated between the two groups. One of these genes was prolactin (*PRL*), which was found to be up-regulated in infanticide animals. *PRL* plays a central role for several behaviours such as feeding, stress, fertility and aggression. High level of PRL have been linked to aggressive and hostile behaviour [81]. *PRL* stimulates the release of dopamine which in return inhibits its release via negative feedback [82]. It is also linked to other important genes related to maternal aggression such as oxytocin and *POU1F1* (POU domain class 1, transcription factor 1). Low level of oxytocin have been linked to aggressive behaviour [83]. *POU1F1* regulates *PRL* expression and was found regulated in this study. Another interesting gene linked to *PRL* found to be regulated in this study is *PRLR* (prolactin receptor). It regulates the oestrogen receptor Esr1 and Esr2 via the Jak2-Stat5 pathway [84] and was found to be down regulated in infanticide samples. Another gene directly related to *PRL* is *DRD2* (dopamine receptor 2) which negatively regulates the expression of *PRL* which in turn decreases dopamine levels. It was found to be down-regulated in this study and has been linked to a number of neuropsychiatric disorders such as schizophrenia, post-traumatic stress disorders and migraine [85]. It is also involved in the protection of oligodendrocytes against oxidative glutamate toxicity and oxygen-glucose depravation injury. Such injuries can trigger abnormalities in oligodendrocytes and have been linked to schizophrenia and mood disorder [86, 87]. Other genes have been linked to dopamine levels, such as *POMC* (pro-opiomelanocortin) and *OPRM1* (micro-opioid receptor 1). The first one is down-regulated for the infanticide group and the second one is up regulated. Interestingly both genes are closely linked, endorphin hormones are produced from the precursor *POMC* and bind to *OPRM1*, up-regulation of *OPRM1* could be in response to the lack of endorphins but if the expression of *OPRM1* is reduced it would result in a reduction in the secretion of dopamine [88]. Another gene found to be up-regulated in the infanticide group is *5HTR2C* and was found in to be in a QTL region in the study by Quilter et al [17]. As discussed in a previous section 1.4.2, it is linked to the regulation of dopamine levels, serotine regulates dopamine via this receptor and decreases in its expression have been linked to a reduction of the symptoms of depression [89]. Therefore the up-regulation of this gene could lead to a decrease in dopamine levels and the development of depression. Furthermore it is also linked to the activation of *POMC* and variation in the *5HTR2C* has been linked to bipolar affective puerperal psychosis [90]. The expression of *5HTR2C* has also been linked to other psychological disease [60, 61, 62, 63, 64, 65]. Another gene found to be regulated is *POU3F3,* interestingly, it is also located near a SNP that has been linked to bipolar disorder [91]. A lot of genes are linked to dopamine expression, and dopamine receptors are known to modulate NMDA (N-methyl-D-aspartate) glutamate receptor mediated function. *GRIN1*, the receptor for glutamate is the major excitatory neurotransmitter and showed differential expression in the porcine arrays. It is involved in learning, memory and some aspect of behaviour and schizophrenia and psychiatric disorder [92, 93]. The study also investigated G proteins and MAP Kinase signalling pathway. The central serotonergic, noradrenergic and dopaminergic system are all acting via protein G (Gi, Gs and Gg) and dysfunction in this system has been linked to depression in human [94]. The Adenylase cyclase inhibiting G alpha protein (*GNAI1* or *Gi*) gene was found to be differentially expressed in this study. It is involved in the MAPK signalling pathway which has been shown to be involved in Alzheimer's Disease [95].

Another pathway that could be involved in maternal infanticide is the Oestrogen pathway. Transthyretin (*TTR*) was found to be differentially expressed, it is a transporter of thyroid hormones and Vitamin A. Its levels are altered in cerebral spinal fluid of psychiatric patients and KO mice have shown reduced depression [96]. Therefore the increase in expression seen in three of the individuals suggests an increase risk of depression. Overexpression of *TTR* correlates with increased levels of oestrogen, which has been linked to increased aggression towards offspring [9].

Several other interesting genes linked to the mitochondrial pathway are differentially regulated in this study. *ATP6*, *ATP8 COX1*, *ND2*, *16S*, *ND4* and *ATP5A1* are found to be differentially expressed in this study on at least one of the various analysis. As the brain has a high aerobic activity and therefore requires a high mitochondrial content and activity, it is therefore more likely to suffer from mitochodrial defect and diseases. Such issues have been shown to play a role in a wide range of brain disorders such as schizophrenia, bipolar disorder, Alzheimer's Disease, epilepsy and Parkinson's disease [97].

1.4.4 Association Study

More recently, porcine infanticide was investigated using association testing in [2]. The study was performed on four different lines of pig: B,C,D and H, consisting of pure breed Landrace LR (B), pure Large White LW (C) and crosses with Duroc (Duroc x LR for D, Duroc x LW for H).

The study used the pig SNP60k array from Illumina [98]. In total 225 animals were investigated in this study and after QC, 210 animals were left with around 50 animals for each line. Different approaches were used to investigate the data, a test of association was performed, a haplotype test using sliding windows was also performed and a haplotype based analysis. Two lines (D and H) had also enough related samples to perform a sib pair analysis. Using various criteria for filtering the results, see [2], several regions of the genome were identified by the different type of tests performed. Using the synteny with the human genome, Quilter et al were able to select regions that have interesting genes, with function related to maternal infanticide or brain function. Some of the regions also contain some of the genes identified in the gene expression study discussed in section 1.4.3 and described in [80], notably *POU1F1* and *GRIN1*. Furthermore some regions identified match regions of the QTL found on chromosome 2, 10a and 10b by [17, 14] and discussed in section 1.4.2.

The most interesting region identified by this study is a large region on SSC3 which is consistent across all the analyses. Two different peaks are presents in this region, one at 23.4MB and one at 31.9MB. The first peak is the top SNP for the single SNP analysis and mapped to a region between *GPR139* and *IQCK*. Interestingly within 1MB of this peak, the corresponding region in human has been found to be involved in bipolar disorder, with half of the female patients in these studies exhibiting post-partum symptoms [99, 100]. There are several potential candidate genes in this region, such as the regions from *XYLT2* to *ABCC1*, *SYT17* to *XYLT1*, *PARN* to *MLK2*, *ERCC4* to *SHISA9*, which have all have be connected to some psychological pathology such as attention deficit hyperactivity disorder (ADHD), alcoholism, Autism, Alzheimer's Disease, conduct disorder and schizophrenia [101, 102, 103, 104, 105]. *GPR139* is an interesting gene, it is a G-protein coupled receptor, important in signal transduction. In the mouse it is expressed in the putamen, medulla and caudate nucleus and also in the thalamus, amygdala and spinal cord but at lower levels [106]. Those areas of the central nervous system are involved in mood, behaviour and locomotion activities. On an evolutionary perspective, *GPR139* has been placed in the same group as somatostatin receptors [107]. Somatostatin plays an important role in the regulation of hormone release, as parturition is an event leading to an important change in the hormonal balance of the mother, meaning this gene could contribute to post-natal behaviour changes.

The second peak at 31.9MB is in a region between *RBFOX1* and *UBN1.* This region is syntenic with a region on the human chromosome 16p13 which has been linked to puerperal psychosis [108] and it is within 2MB of a region linked to bipolar disorder [109]. RFBOX1 binds to the C-terminus of ataxin-2 and may contribute to the restricted pathology of spinocerebellar ataxia type 2 (*SCA2*). Ataxin-2 is the product of the *SCA2* gene and causes familial neurodegenerative disease [110]. RFBOX1 is also involved in movement related adverse psychotic effect for conduct disorder [111] and autism [112] and was a potential candidate for ADHD [113]. The syntenic region in human to *RFBOX1* has SNPs that are associated with bipolar disorder, autism, Alzheimer's disease, ADHD and hyperactivity conduct disorder [114, 115]. It is also linked to *GRIN1* as it is a modulator of neurally expressed genes. As discussed before, see section 1.4.3, *GRIN1* was found to be differentially regulated in [80] and is implicated in a number of psychiatric disorders and diseases [93, 92]. The second gene in this region, *UBN1*, is in a region that has been investigated for its potential involvement in autism in human [116]. Those two regions are not in LD and therefore form two different blocks.

Another interesting region identified is on SSC4 which is syntenic for a human region on chromosome 16 which has been linked to puerperal psychosis [108]. Some segments on SSC4 reached genome wide significance, the region extends from *FAM135B* to *KHDRBS3* and SNPs in the intergenic region have been associated with Parkinson's disease, neurotic disorders and cognition in human [117, 118, 119]. *PTK2* is in this region and is associated with autism in human [120]. *KCNK9* is also in this region and a member of the superfamily of potassium channel proteins containing pore forming P domains. It is highly expressed in the cerebellum and is imprinted in human and mouse fetal brain. It has preferential expression from the maternal allele. It is mutated in a maternally transmitted genomic imprinting syndrome of mental retardation [121], which makes it an interesting candidate as it has been show that the heritability of maternal infanticide is higher from dam to daughter than sire to daughter [4, 3].

A region on SSC9 was identified as being of potential interest after the permutation analysis. It matches by synteny a region on human chromosome 1 containing *METTL13.* This gene has been found to be linked to increased susceptibility to postpartum mood symptoms [122]. *HMCN1* has also been involved with those symptoms and is just outside the region identified.

A region identified on SSC15 reached genome wide significance and has a few interesting candidates. *PAX3* is found in this region and is a critical gene in fetal development, more specifically in neural development as mutations in this gene lead to spina bifida and exencephaly [123]. Mutation of *PAX3* in human are also linked to the central nervous system, they have been associated with the craniofacial-deafness-hand syndrome [124]. Another interesting gene is *EPH4*, a member of the epharin receptor family which is involved in mediating developmental events in the nervous system [125]. It seems also to contribute to Alzeihmer's Disease, it has been associated with density, volume and cortical thickness of the hippocampus [126].

1.4.5 Table summary

Table 1.7 summarises the various genes and the pathology or functions in relation to maternal aggression found in the studies discussed in section 1.4.2, 1.4.3 and 1.4.4.

1.5 Microarray technology

In general for an association study to be successful it must have a large number of markers and participants compared to linkage studies. The discovery of frequent Single Nucleotide Polymorphisms (SNP) during the sequencing of the human genome in the early 2000s unlocked the possibility of performing this type of analysis on cohorts of individuals [127]. This was combined with the advance in technology for typing DNA and the invention of the DNA microarray [128] which allows GWAS data to be collected at a reasonable cost. The DNA microarray technology is a significant advance in increasing the throughput of genomic studies and can be used for a wide range of applications: gene expression, methylation, copy number variation (array Comparative Genomic Hybridisation or array CGH), chromatin modification (Chromatin Immuno Precipitation or ChIP) and genotyping profiling being the major ones. The principle of microarrays is simple, a probe fixed on the array (a solid support) hybridises in solution with a target with usually one dye (fluorochrome) attached to it. The dimer formed between the target and the probe will emit a signal that is used to quantify or perform a "present/absent" call on the dimer. In order to type a sample, the DNA (or RNA) has to be fragmented, amplified by Polymerase Chain Reaction (PCR), which converts the sample to cDNA in the case of RNA, and then is hybridised to the probe attached to the surface of the array. Those probe sequences can be tailored to the specific application, using sequences complementary to a given gene for gene expression, having two or more probes to type specific SNPs or the methylation status of a base. Once hybridised the array is washed to remove any non bound DNA and then the array is treated to add or enhance the fluorescent signal. The array is then scanned using confocal lasers to excite the fluorochrome and a high resolution camera to image them.

Microarrays can use a single fluorochrome (single channel microarrays) or two different fluorochromes (dual channel microarrays). The latter uses fluorochromes that when excited will emit light at well separated wave length. These arrays are used for competitive hybridisation in order to investigate differential expression analysis or perform array CGH. Dual channel microarrays were the preferred choice when the technology was progressing as it helps to reduce the potential bias coming from the use of multiple arrays to process many samples. Usually the design had to include dye swapping in

Chromosome	Genes involved	Pathology/Gene function	Study type	Study
1	*OPRM1, GRIN1*	Dopamine regulation, depression, schizophrenia	GX,AS	[80, 2]
2	*COX7C, YY1, GR*	Bipolar disorder, schizophrenia, neural degeneration, depression, panic disorder agoraphobia	Linkage,AS	[17, 2, 14]
3	*POMC, POU3F3,* *GPR139 to IQCK, RBFOX1 to UBN1*	Dopamine regulation, depression, bipolar disorder, ADHD, alcoholism, autism, Alzheimer disease, schizophrenia, puerperal psychosis, neurodegenerative disease, conduct disorder	GX, AS	[80, 2]
4	*FAM135B to KHDRBS3*	Puerperal psychosis, Parkinson's disease, neurotic disorders, cognition, mental retardation	AS	[2]
6	*DPB,TTR*	Audiogenic epilepsies, aggression	Linkage, GX	[80, 14]
7	*PRL*	Aggression and hostile behaviour	GX	[80]
9	*DRD2, GNAI1, METTL13, HMCN1*	schizophrenia, post-traumatic, movement and stress disorder, migraine, Alzheimer Disease, post-partum mood symptoms	GX, AS	[80, 2]
10	*PTPRC (CD45)*	Alzheimer disease, Down's syndrome	Linkage,AS	[17, 2]
13	*POU1F1*	Regulate PRL	GX, AS	[80, 2]
14	*ADRA2A*	Schizophrenia	Linkage	[14]
15	*GPM6A, PAX3*	Schizophrenia, spina bifida, exencephaly, craniofacial-deafness-hand syndrome, Alzheimer's Disease	Linkage, AS	[14, 2]
16	*PRLR*	Oestrogen regulation	GX	[80]
X	*STS, GLRA2,PGRMC1, 5HTR2C*	risk deficit hyperactivity disorder, autism, social communication deficits, aggressive behaviour, schizophrenia, infanticide, depression, suicide, alcoholism, anorexia.	Linkage, GX	[17, 80, 14]

Table 1.7: Table summary of all the loci and gene found in the different studies. AS stands for association study and GX for gene expression study

order to counter the bias due to the difference of signal strength coming from a specific dye [129]. The technology is prone to a number of technical biases coming from the production of the array (print tip variations for example), the extraction of the RNA/DNA and the preparation of the arrays. In order to correct these biases, normalisation methods were developed [130]. These methods were designed to remove technical biases from the data generated by microarrays. Technical biases are defined as systematic variation in signal, the origin of which comes from the technology used. The data obtained from the scanner is called "raw data" and needs to be processed in order to produce "normalised data", which should be free of technical biases. Various normalisation methods were used at first, as the technical biases were important, nowadays the data are transformed on a log_2scale (to compress its scale) and normalised using a quantile normalisation. The resulting data will all share the same distribution of signal which will allow better statistical analysis. The evolution of the normalisations methods followed the improvement of the methods used to produce arrays.

At first, most groups were producing their own microarrays or had access to a dedicated facility to do so. There were a substantial number of methods used to produce the microarrays[129]. One of the most widely produced microarrays in laboratories were spotted arrays, where the probes were laid onto a slide using inkjet or contact printing technology[131]. Other more sophisticated methods use in-situ synthesis of oligomucleotide probes. As the technology improved, the production moved from laboratory to commercial companies, which developed more robust methods, thus improving reliability [132, 133]. While the first microarrays had a few hundred to a few thousands of probes on them, this quickly increased with the advance in technology triggered by the involvement of biotechnology companies. These improvements resulted in more robust, reproducible and cheaper arrays while also leading to an increase in the number of probes present on the arrays and in parallel with an increase in the number of samples being processed. This represented a major breakthrough for GWAS as the sample sizes and number of markers needed for successful studies are large. The sample size needed to achieve good power for GWAS is linked to the number of markers studied. The larger the sets of SNPs, the more samples are needed in order to achieve good power [42]. The markers are chosen in order to represent LD blocks as defined below and thanks to the advances in technology, typing thousands of SNPs on large cohorts became more viable, transforming GWAS into the major approach in the study of genetically linked diseases [134].

In order to design good genotyping microarrays for GWAS, the selection of appropriate markers is crucial. Ideally the markers should be selected to represent the genome as finely as possible in order to identify any association between genotype and phenotype. The selection of markers is done using LD to select 'tag' SNPs. As discussed before LD is characterised by the joint transmission of alleles, and can be measured in various ways [135]. One measure is $D_{AB} = p_{AB} - p_A p_B$ which is the frequency of the heterozygote allele minus the product of the frequency of both individual alleles. If the loci are in linkage equilibrium then $p_{AB} = p_A p_B$ and $D = 0$. This measure is not the most useful measure to compare between pairs of markers as it is tied to their allele frequencies. D can be normalised by defining D_{max} as the smaller value between $p_A(1-p_B)$ and $p_B(1-p_A)$ when $D_{AB} > 0$, and the less negative value between p_Ap_B and $(1-p_A)(1-p_B)$ if $D_{AB} < 0$ be then we can define $D'_{AB} = D_{AB}/D_{max}$ as suggested in [20]. The normalised D_{AB} can be used to compare pairs of markers. Another measure for LD is $r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$ which is the Pearson correlation coefficient between the markers. The various ways of measuring the LD can be used to define LD for regions of the genome using known markers. It can be used to define blocks of LD or haplotype blocks of markers (markers transmitted together). Defining those blocks allows the selection of 'tag' SNPs for the construction of microarrays to test for association of markers to disease contributing loci using case/control studies [136]. Tag SNPs are choosen to represent regions of the genome, they allow us to indirectly type other markers in the LD block they belong to as those markers are transmitted together because of the linkage disequilibrium. It is therefore possible to infer the allele of other SNPs in this region using the tag SNPs.

The array used for this study is the porcine SNP 60k array by Illumina, the design process is described in details in [98]. Briefly, libraries of pool of animal from five different pig breeds (Duroc, Pietrain, Landrace, Large White and Wild boar) were sequenced using short read (36bp) and longer read on the Roche 454 platform. The reads obtained were aligned against the pre-release version 7 of the pig genome for SNP discovery. After pruning and selection, the final number of SNPs selected for the array was 64,232.

1.6 Deep Sequencing technology

Once regions of interest have been identified by using tag SNPs or microsatellites, they can be studied in more detail using deep sequencing. This methodology is used to sequence the full or part of the genome in order to get more information about the surrounding region of the significant SNPs. Following up on a GWAS to study the regions discovered with a deep sequencing approach has proven useful to identify causal variants for several diseases [137, 138, 139, 140]. Deep sequencing allows a detailed investigation of the sequence of the genome and therefore refine any leads given by the GWAS.

1.6.1 Deep sequencing methodology

Over the last 50 years the methods used to sequence DNA have greatly advanced, especially since the completion of the first draft of the human genome. Several methods have been developed in order to perform deep sequencing of the DNA [141, 142]. One method was originally developed by a company called Solexa and then bought by Illumina and became the most widely used sequencing technology [142]. This technology performs sequencing by reversible terminator chemistry [143]. The principle starts by first shearing the DNA in small fragments. The fragments are ligated with adaptor sequences specific to the Illumina platform. Each end of the fragment has a specific adaptor which also contains the sequencing primer. The fragment can also be labelled with specific indices sequences (also called barcode), which can be used to track the samples they originated from. Fragments from different samples can then be pooled together, usually to save on the cost of sequencing . The fragments once ready to be sequenced (pooled or not) are called a sequencing library. The library is then loaded on the instrument. The DNA fragments will hybridise at one of their ends with the adaptors present on the flow cell used for the sequencing. The flow cell consists of a glass slide with one or more channel (also called lanes) that have complementary sequences to the adaptors used during the library preparation. Once hybridized the fragments will undergo an amplification step also called “bridge PCR”, see figures 1.2, 1.3, 1.4. The free end of the fragment will bind to a nearby free adaptor sequence fixed on the surface of the flow cell and form a “bridge”. This “bridge” is then used to perform an amplification step resulting in the clonal copy of the original fragment: the complementary sequence is removed after the amplification step. The amplification step is repeated several time to generate a “cluster” of clonal copies of the same fragment. The number of clusters being sequenced will depend on the sequencer used. Once the amplification step completed, the clusters will be sequenced by reversible terminator technology, figure 1.5. The technology works by using the primer sequence present on the adaptor sequence to start a polymerase reaction. Primers are flushed on the flow cell to start the reaction. Then the sequencing is started and consists of several cycles. A cycle starts with the four nucleotides being flushed onto the flow cell. The nucleotides are labelled with fluorochromes and have a blocking group (3’-OH) that will result in the addition of only one base during each cycle. Once the reaction is completed for that cycle, the remaining nucleotides in solution are washed away and the instrument will image the flow cell surface, taking a picture of the cluster colour for each cycle. As all the fragment forming a cluster are of the same composition, the fluorescent signal emitted will be the same for all the fragments forming a cluster. After the scanning step the blocking group is removed so that a new cycle can start. Once the sequencing is completed, the sequence generated for a cluster of the flow cell is called a read. The length of the read will be based on the number of cycles performed, which will be determined by the user prior to sequencing. At the present the maximum sequencing length (also called read length) for one end of the fragment is 250 nucleotide for the MiSeq platform. Sequencing can be done from one end of the fragment, called single end sequencing, or from both end, called paired end sequencing, which is done by paired end “turn around” chemistry. The strands are flipped on the flow cell by performing several cycles of bridge amplification, therefore allowing sequencing of the other end of the fragment.

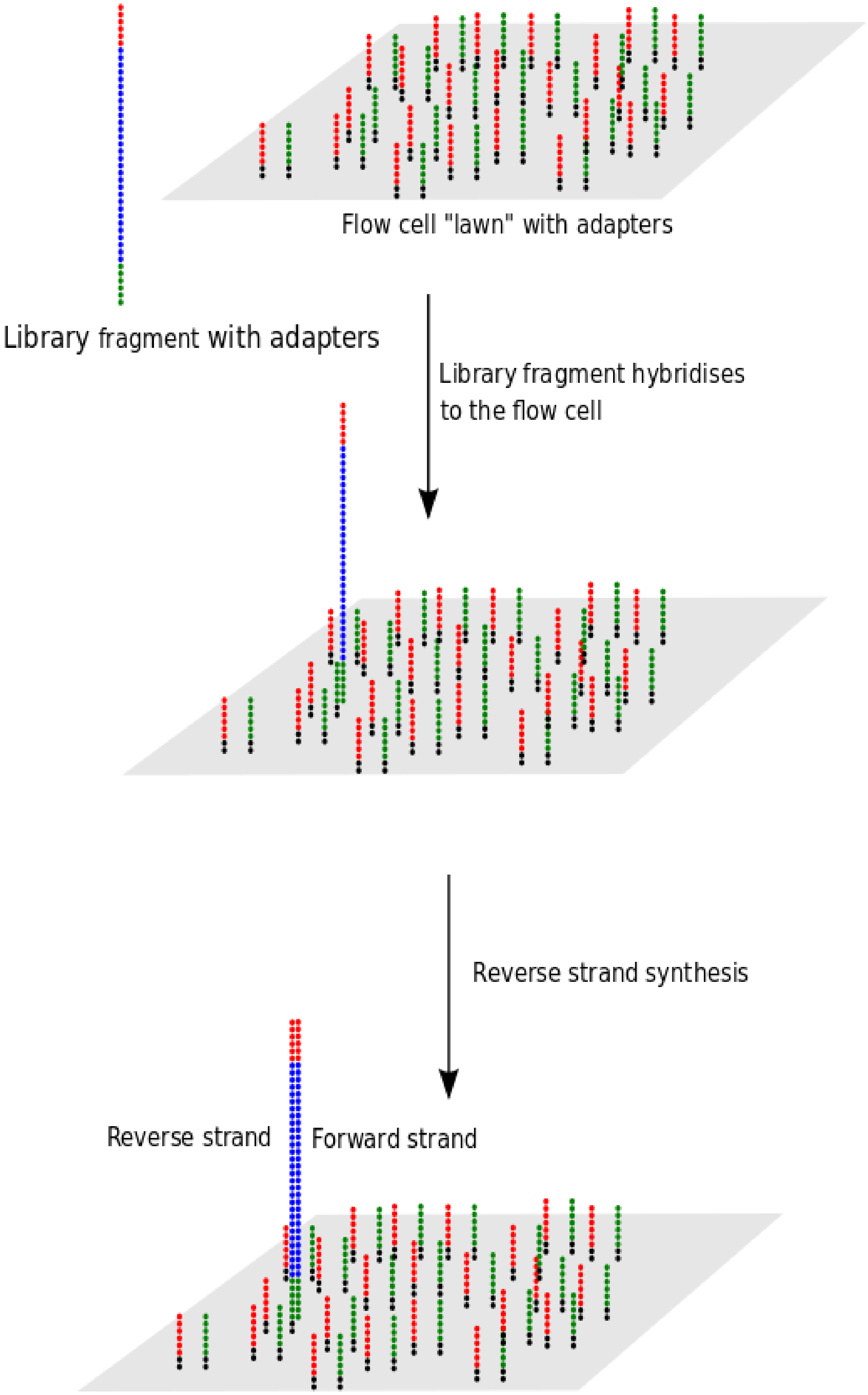

Flow cell "lawn" with adapters
Library fragment with adapters
Library fragment hybridises
to the flow cell
Reverse strand synthesis
Reverse strand
Forward strand

Figure 1.2: Bridge PCR amplification first phase, reverse strand synthesis

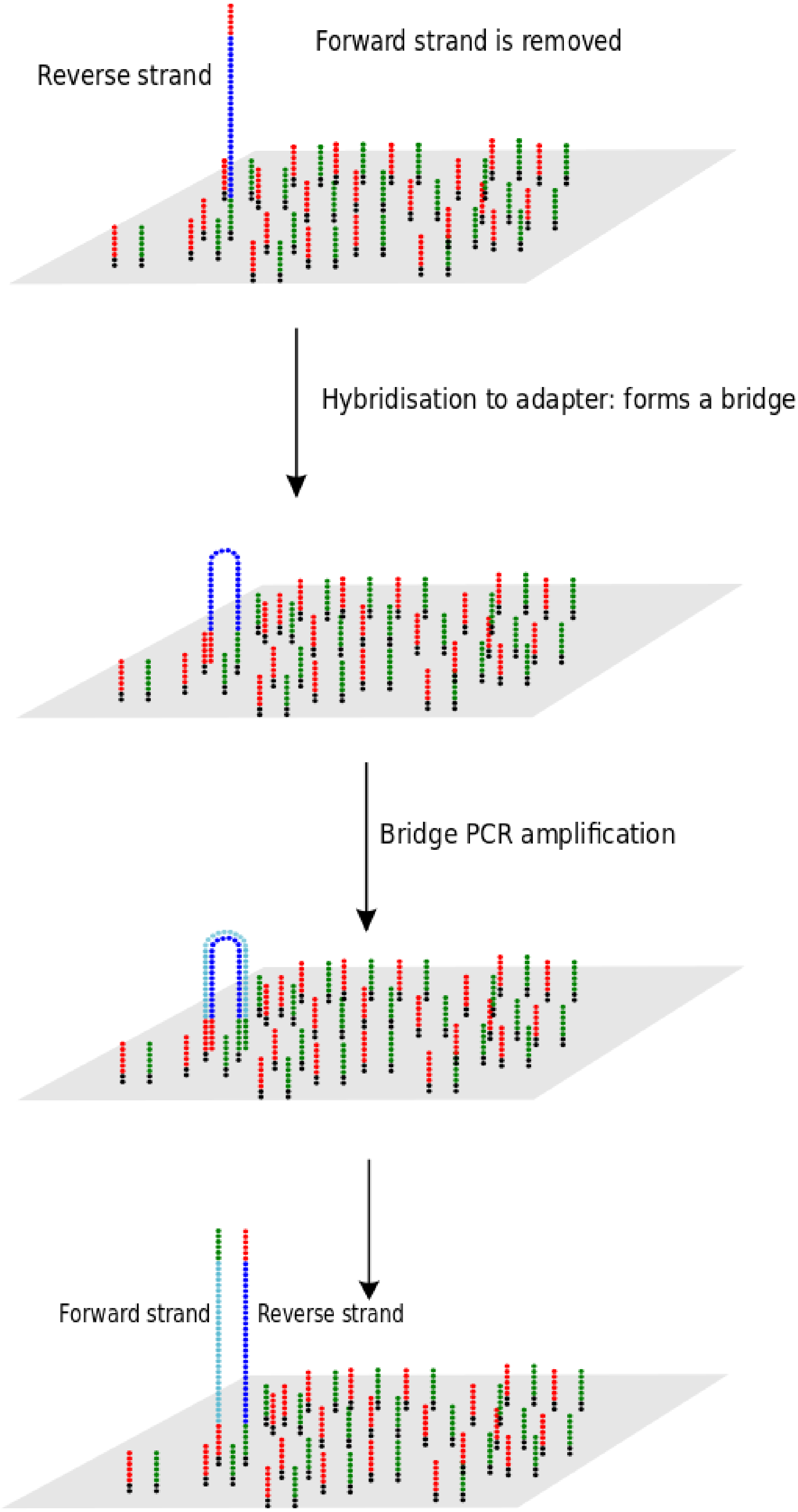


Figure 1.3: Bridge PCR amplification second phase, bridge PCR

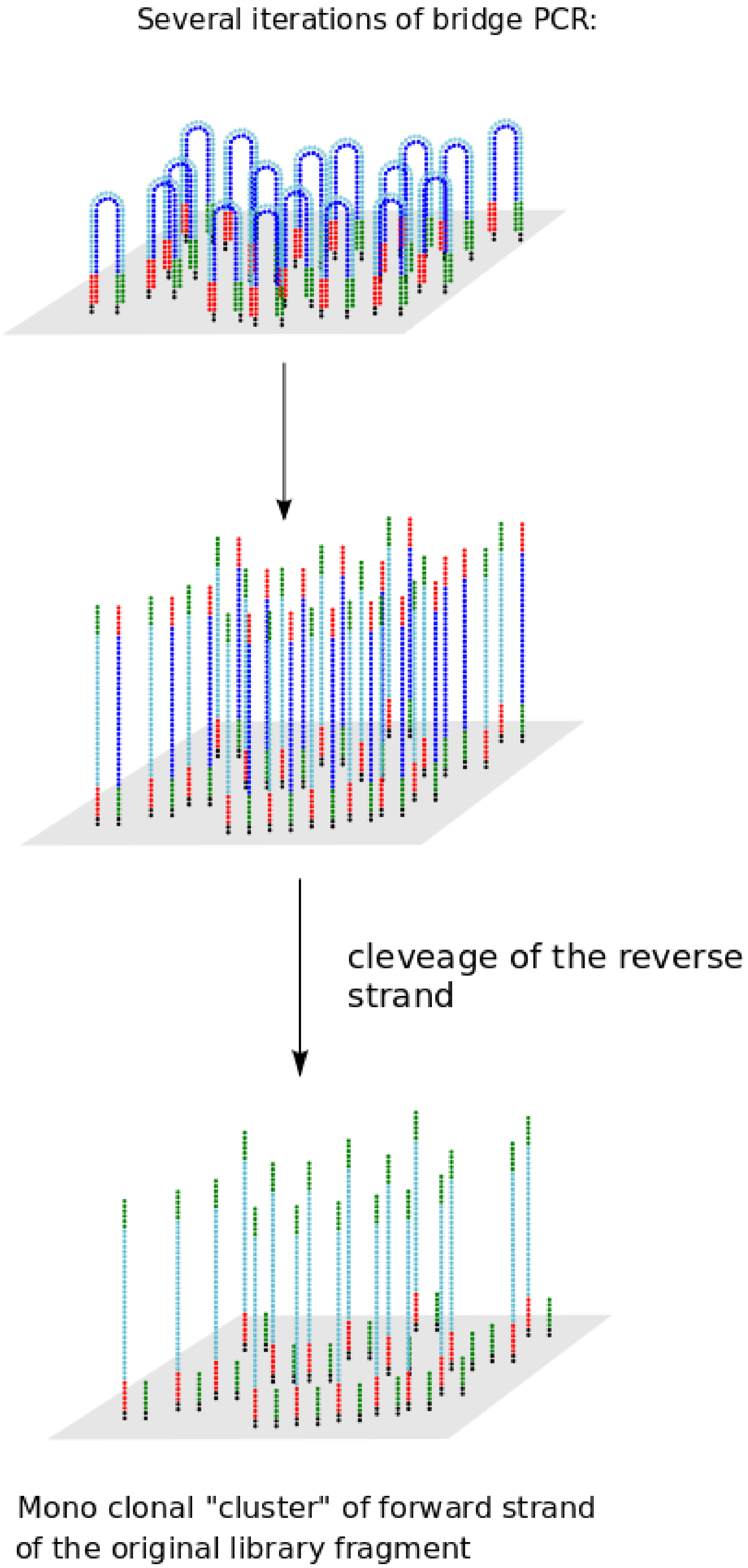

Several iterations of bridge PCR:
cleveage of the reverse strand
Mono clonal "cluster" of forward strand of the original library fragment

Figure 1.4: Brige PCR amplification third phase, mon-clonal amplification

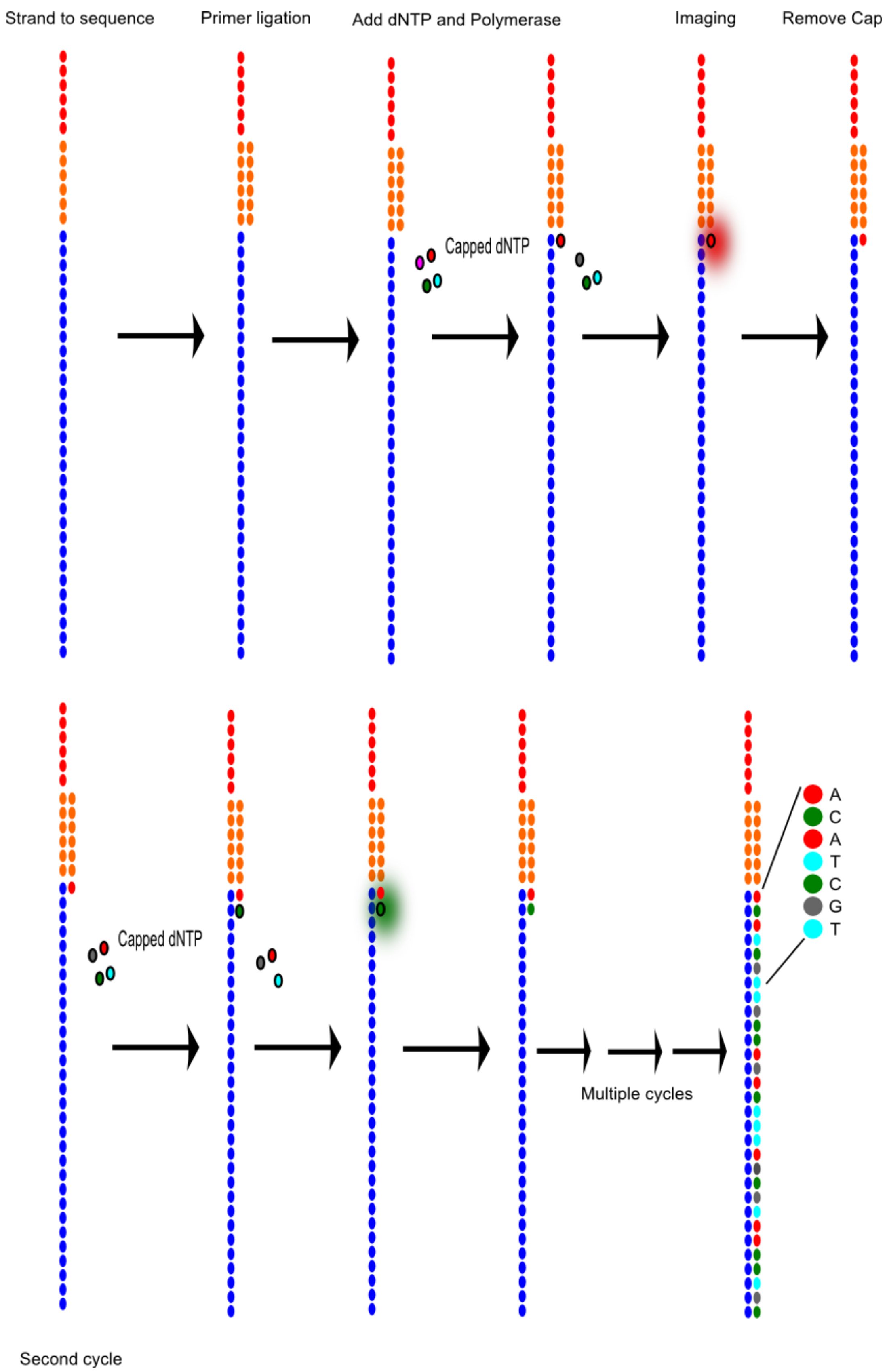


Figure 1.5: Sequencing by synthesis, strand sequencing using reverse terminator.

1.6.2 Targeted sequencing

While a full genome sequencing approach will give the largest amount of information, it is not always the most cost effective way to identify variants in a region of interest as a large portion of the sequencing data will not hold any useful information. New approaches have been developed in order to sequence only part of the genome, called targeted sequencing, sequence enrichment or sequence capture. Manufacturers have designed pre made panels for various applications but most of those are available for human genomes only. For non model organisms, once the regions of interest have been identified it is possible to design a custom capture kit to target the regions of interest that have been identified using an association or linkage approach. Several companies are offering the design of custom capture panels, Roche Nimblegen (SeqCap), Agilent (SureSelect) and Illumina (Nextera Rapid Capture). Agilent SureSelect uses long, 120-mer, biotinylated cRNA baits to capture regions of interest in order to enrich them from a genomic fragment library. The processe is described in figure 1.6: it starts by the shearing of the genomic DNA (gDNA) and standard library preparation. Then the sample library is hybridised with the baits and the hybridisation product is pulled down using magnetic streptavidin coated beads. The selected library is then amplified and put on the sequencer. Nimblegen SeqCap uses the same principle as the SureSelect method, as seen in figure 1.7. Illumina Nextera Rapid capture also follow a similar protocol but instead of using traditional library preparation methods, it uses the Nextera library preparation protocol which take less time than traditional methods as it uses transposons to cut and add the sequencing adaptors to the gDNA which make the process a lot faster. Once this step is completed the sequence capture is carried out in the same fashion as the other method, using biotinylated probes and streptavidin coated beads.

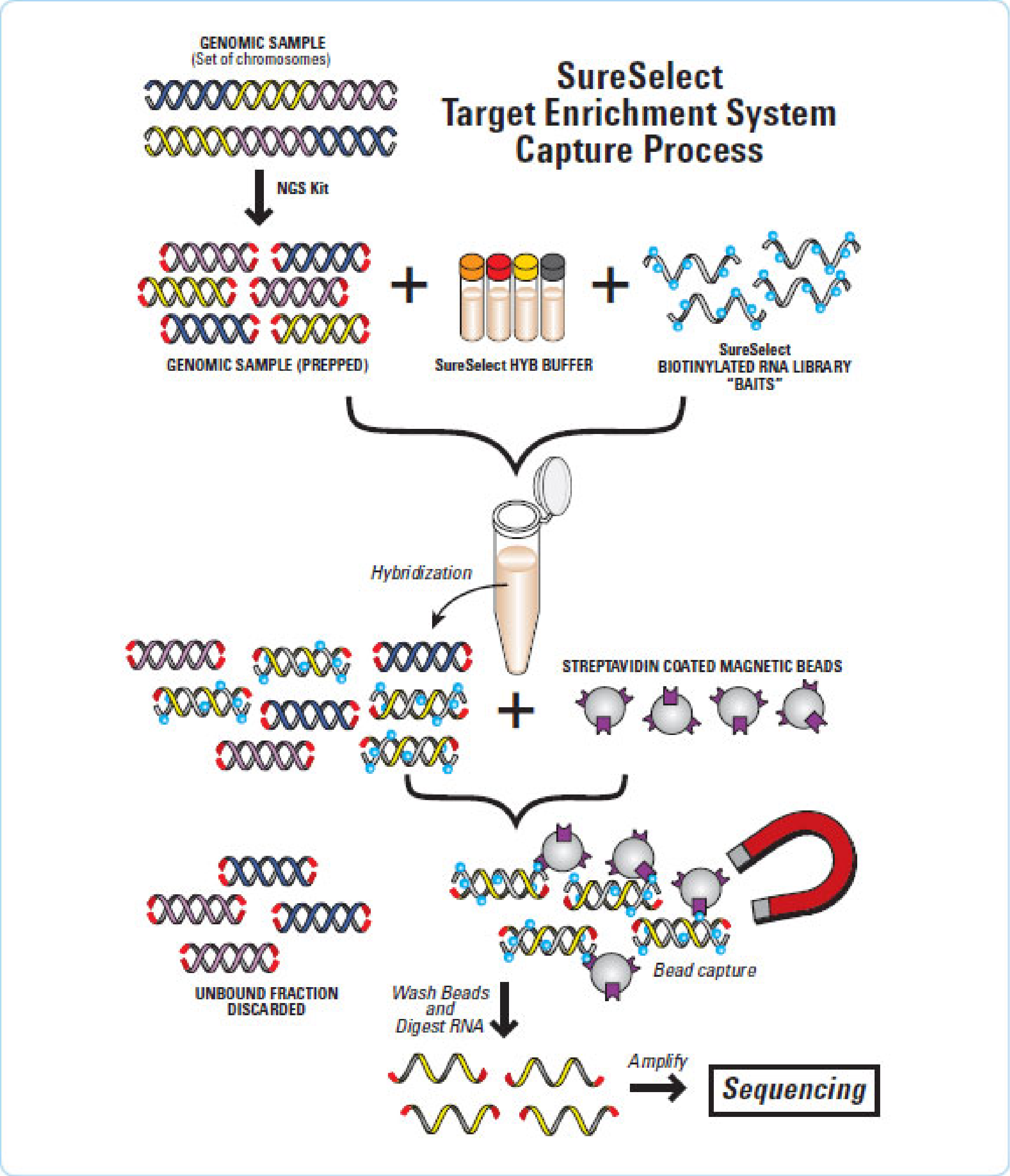


Figure 1.6: Agilent SureSelect sequence capture. Source: Agilent website (www.genomics.agilent.com)

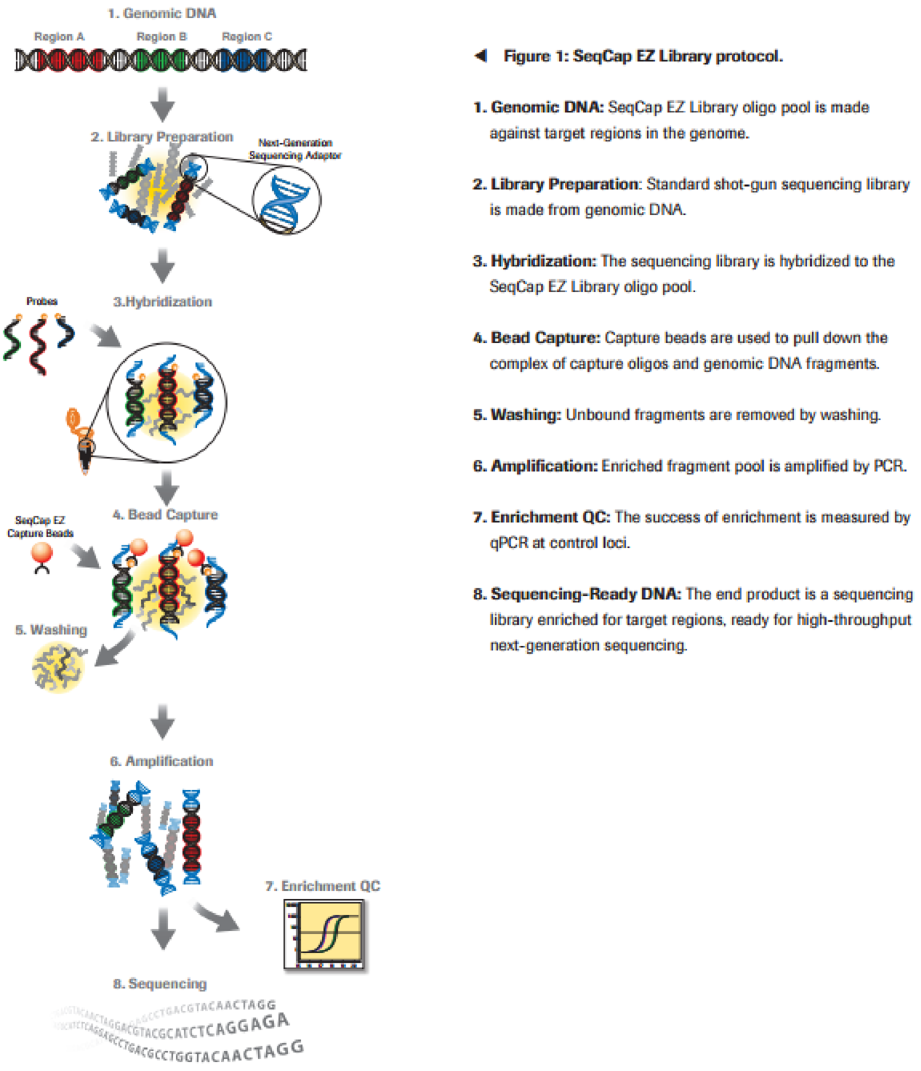


Figure 1: SeqCap EZ Library protocol.

1. **Genomic DNA:** SeqCap EZ Library oligo pool is made against target regions in the genome.
2. **Library Preparation**: Standard shot-gun sequencing library is made from genomic DNA.
3. **Hybridization:** The sequencing library is hybridized to the SeqCap EZ Library oligo pool.
4. **Bead Capture:** Capture beads are used to pull down the complex of capture oligos and genomic DNA fragments.
5. **Washing:** Unbound fragments are removed by washing.
6. **Amplification:** Enriched fragment pool is amplified by PCR.
7. **Enrichment QC:** The success of enrichment is measured by qPCR at control loci.
8. **Sequencing-Ready DNA:** The end product is a sequencing library enriched for target regions, ready for high-throughput next-generation sequencing.

Figure 1.7: Nimblegen SeqCap sequence capture. Source: SeqCap Brochure 2015

1.6.3 Sequencing mapping

Once the library has been sequenced, the data generated are mapped to the reference genome and analysed using an algorithm designed to call the variant. The variant identified can be evaluated for potential impact on the gene expression or product of critical genes. Because the DNA consists of only four letters, it is easy to index the reference and query sequence in order to perform fast alignments and find a match between the read sequence and the reference. The software used to align the reads sequenced is called Burrows Wheeler Aligner (BWA) [144] which uses Burrows Wheeler Transform (BWT). BWT is useful to speed up the search for a match between the reads and the reference sequence. The reference is transformed using the Burrow Wheeler transform as illustrated in figure 1.8. This step is also called: the indexing of the genome. The reference will be broken up into thousands of BWT words during the indexing. The transformed reference, called the "indexed reference", is used to speed up the search of patterns (queries) in it, in this case the reads.

The BWT has two important properties. First, by sorting the BWT of the word by lexicographical order we can retrieve the first column of the BWT matrix as explain in figure 1.9. This creates an incomplete Burrows Wheeler matrix that can be used instead of the full matrix for pattern searching. Second, another property of the BWT is that the first occurrence of character 'X' in the last column is the same character as the first occurrence of 'X' in the first column. This is easy to see this when looking at the Burrows Wheeler matrix in figure 1.8. With those two properties it is possible to search a pattern without the full matrix, using only the first and last columns1.9, it is possible to search for patterns. Take the pattern 'ANA' which occurs twice in the world BANANA. The search can be performed with the following simple manual steps:

1. Look for the occurrence of the first letter 'A' in the last column, there are 3 such occurrences, in position 1, 6 and 7
2. Because of the cyclical permutation used to generate the BWT, the letter after 'A' in the original word will be in the first column of the BWT matrix
3. Thanks to the property of the BWT we can find the first column from the BWT word by sorting it by lexicographical order.
4. In order to identify the correct occurrence of 'A' in the first column, we use the property previously described, the i^{th} occurrence of a letter in the last column is the same as the i^{th} occurrence of the same letter in the first column. Therefore we know that the first occurrence of it will be followed by '$' which does not match 'N', the next two occurrences do match our pattern. The pattern 'N' in the first columna and 'A' in the last matches row 6 and 7 of the BWT matrix.
5. The second letter for our search is N, which corresponds to 2 of the 3 occurrences of 'A' identified in the previous steps.
6. By checking the 'N' in the last column and looking at the first column we know that both are followed by 'A' so our search is over, but if we were to continue we would apply the same principles.
7. The matching rows of the BWT matrix are the second and third occurrences of A in the first column, so line 3 and 4.

However in order to use this method more efficiently, a few mathematical transformations allow the automation of the search for large scale pattern mapping.

The method using a Last to Front (LF) mapping table can be used to perform a search for matching characters:

1. Calculate the C array for the BWT of the word to reconstitute, C(i) corresponding to the number of characters lexicographically smaller in the BWT word, for example for B there are '$' and 3 times 'A', so C(B)=4. See table 1.8

2. Perform the LF mapping, using the following formula $LF(i_j) = C(i) + n_i$ with i_j being the j^{th} occurrence of the character i and n_i being the number of occurrence of the character$i$ up to and including i_j. See table 1.9.
3. Then get the first column of the BWT matrix by sorting the BWT by lexicographical order : $AAABNN
4. Seaching for 'ANA' we start again by the last character 'A'
5. The As have as LF value 2, 3 and 4 in the table 1.10 when checking the top row (BWT).
6. Checking the column 2, 3 and 4 of the first row for the next letter 'N', there are two matches for column 2 and 3 with LF numbers of 6 and 7.
7. Checking column 6 and 7 for the next letter 'A', both match, therefore the word 'ANA' is found twice.
8. The LF to front value is 3 and 4, which indicate that the pattern start on the first column of the BWT matrix at row 3 and 4.

This method is also useful to exclude mismatch or partial match data, looking for ABA in BANANA:

1. Starting by the last letter 'A', the LF value are 2, 3 and 4.
2. The next letter is 'B', checking column 2, 3 and 4, B is present in column 4 with a LF value of 5.
3. The next letter is 'A', however the letter in column 5 is '$' which does not match therefore the pattern 'ABA' is not present in our word.

The next question when one or several matches are found is to know where the match is located in the original word. Going back to the pattern 'ANA' which had 2 occurrences in our our word, using a "walk back" approach and counting the number of steps (or offsets) we can get the position of the match in the original word:

1. As determined before the 2 rows which start with the 'ANA' pattern are rows 3 and 4.
2. Starting with row 3, the letter for the last column is N with a LF of 7.
3. The 'N' at row 7 has a 'A' in the last column with a LF value of 4, the offset is set to 1.
4. The 'A' on the 4th row of the matrix as a 'B ' in the last column with a LF value of 5, offset is set to 2.
5. The 'B' on the 5th column of the matrix has an LF value of 1 with the character '$' in the last column, offsets is set to 3.

6. As '$' is the top of the matrix we have completed or walk back to get the position of the pattern in our original word

7. Therefore the first 'ANA' pattern start 3 character after the first in BANANA (so the 'A' which is the 4th letter in the word is the starting position)

8. For the second pattern we start at row 4, which as a 'B' in the last column with a value of 5.

9. The 'B' on the 5th column of the matrix has an LF value of 1 with the character '$' in the last column, offsets is set to 1.

10. As '$' is the top of the matrix we have completed or walk back to get the position of the pattern in our original word

11. Therefore the second 'ANA' found starts 1 character after the first in BANANA.

When working with millions of reads and longer words and patterns, this method is too slow. One alternative is to store the offset value for each line into a suffix array of the BWT matrix along with the BWT, but this is also ineffective as the size of the offset matrix will be very big, 12G for the human genome [144], however new methods allow the size of the suffix array to be reduced to more manageable and efficient sizes [145].

The BWT word can also be reversed to the original word using the following method:

1. We start by the last letter of the word. Because we added '$' to the end word of the word, and that the BWT is sorted, we know that the last letter is first one in the BWT (first column to last column of the BW matrix). In this case the last letter is A and has a LF of 2.

2. Using the LF value as a reference for the column to get the next letter: the second column of our table 1.10 and has the BWT value, 'N'. The second to last letter is N, so the word finished by 'NA'.

3. Again using the LF value as reference, in the second column the value is 6, the 6^{th} column in the table has a BWT value of A and a LF value of 3. Our word finishes with 'ANA'.

4. The third column has a BWT of N and a LF value of 7. Our word finishes by 'NANA'.

5. The seventh column has a BWT of A and a LF value of 4, our word is now 'ANANA'

6. The fourth column has a BWT value of B and a LF value of 5, our word is now 'BANANA'

7. The fifth column matches the special character '$', therefore the word from our previous step is the original word.

The examples above are simple and real data might only have partial matches to the reference genome, sequencing technology is not a hundred percent accurate at the moment, therefore it is likely that some reads will have erroneous bases. In case of inexact matching, the algorithm can introduce a random letter and try to continue the matching protocol in order to find the closest match possible. The advantage of the BWT is that the C array and the LF mapping table can be calculated on the fly, only the BWT of the original reference needs to be stored. Therefore prior to mapping the reads, the reference genome needs to be converted using the Burrows Wheeler Transformation. A suffix array is also generated and will be used to get the position of the matches in the reference. The generation of both files will only be performed once and can be used anytime new data needs to be mapped to the reference.

This is the basis for Burrows Wheeler Aligners, more recent implementation of the aligners added some improvement and heuristics enhancements in order to speed the search of patterns but the base principle use for the mapping stays the same. For this study, deep sequencing technology were used on intervals of interest identified using the family based association test and parent of origin test. In addition the interval already identified from a previous study [2] were used to design a capture panel. In total three capture panels were used in order to study the allele frequency of pools of control and aggressive animals to look for markers that could be used to identify susceptibility to infanticide in animals.

Character i	$	A	B	N
C(i)	0	1	4	5

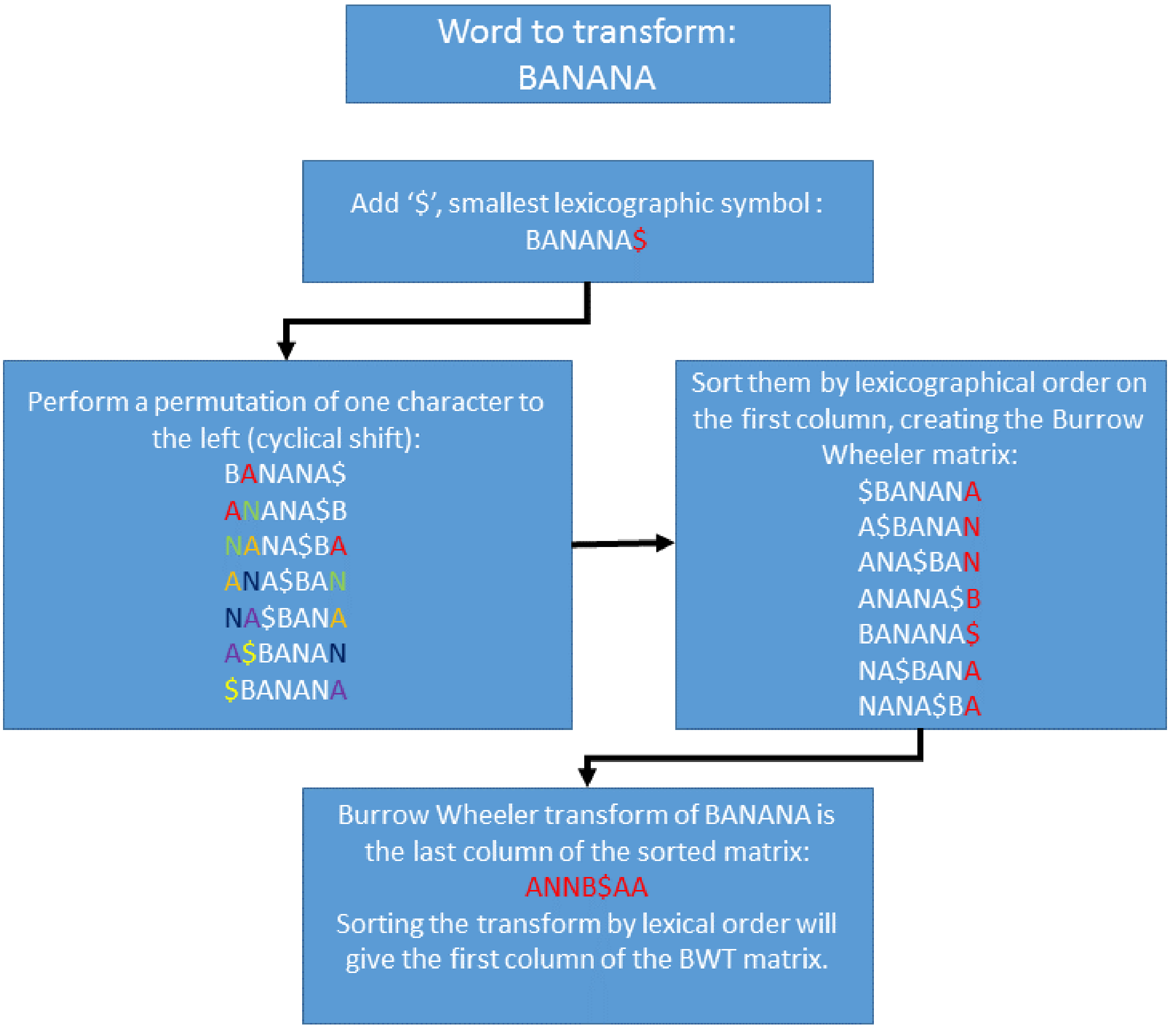
Table 1.8: C array for the BWT transform ANNB$AA . C(i) corresponds to the number of character lexicographically smaller in the BWT word, for example for B there are '$' and 3 times 'A', C(B)=4.

Character i	A	N	N	B	$	A	A
$C(i)$	1	5	5	4	0	1	1
index n_j	1	1	2	1	1	2	3
$LF(i)$	2	6	7	5	1	3	4

Table 1.9: LF mapping for ANNB$AA

BWT	A	N	N	B	$	A	A
First column of BWT matrix	$	A	A	A	B	N	N
$LF(i)$	2	6	7	5	1	3	4

Table 1.10: Table of the BWT, LF values, used to reconstruct the original word



Word to transform:
BANANA

Add '$', smallest lexicographic symbol :
BANANA$

Perform a permutation of one character to the left (cyclical shift):
BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA

Sort them by lexicographical order on the first column, creating the Burrow Wheeler matrix:
$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

Burrow Wheeler transform of BANANA is the last column of the sorted matrix:
ANNB$AA
Sorting the transform by lexical order will give the first column of the BWT matrix.

Figure 1.8: Burrow Wheeler transform

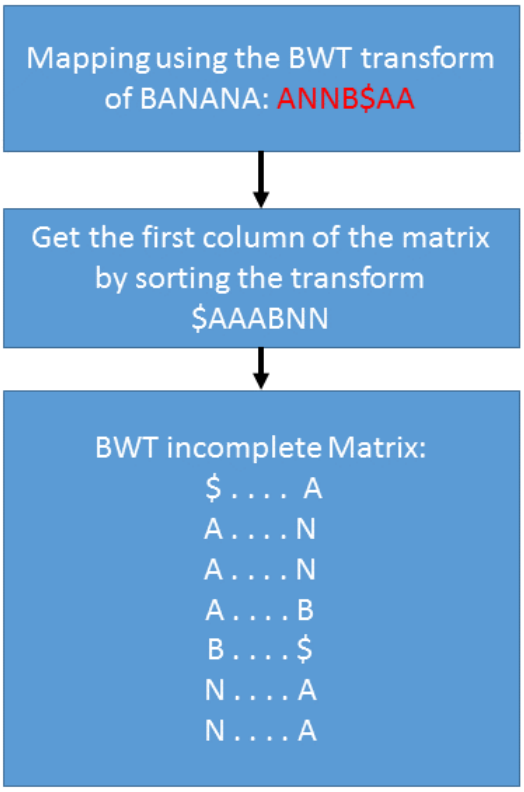

Mapping using the BWT transform of BANANA: ANNB$AA
Get the first column of the matrix by sorting the transform $AAABNN
BWT incomplete Matrix:
$ . . . . A
A . . . . N
A . . . . N
A . . . . B
B . . . . $
N . . . . A
N . . . . A

Figure 1.9: Partial reconstruction of the BWT matrix for pattern searching

1.6.4 Sequence variant calling: unified genotype caller

Once the data have been mapped the next step is to find which bases are different between the reference genome and our capture set in order to identify variants that might be linked to our phenotype of interest. Due to the nature of the sequencing data, looking for base variation is not a simple task, differences can come from a biological source but might also be caused by technical issues arising from the sequencing technology. The error rate for sequencing is not constant and will depend on several technical aspects of the sequencing technology. For Illumina technology the main contributor is the density of cluster on the flow cell. The typical sequencing quality of Illumina generated sequences is expected to have at least 75% of bases above a Phred score of 30 (Q30), which means less than 1 in 1000 error, resulting in a base call accuracy of 99.9%. While this is a very reasonable error rate, it has to be put in perspective with the high amount of data generated by the sequencer. The NextSeq can output between 25 to 120 Gigabases in one run. Furthermore the quality of the sequencing is not even across the length of the read, in general the bases at the beginning of the reads are of high quality and the quality will degrade as the sequencing progress, with the end of the read being of lower quality [146]. This is caused by the degradation of the chemistry as the sequencing progress. The Illumina reported error rate is also not always accurate and the real error rate is often higher [146, 147]. Inaccurate error rate leading to higher than reported rate are problematic as they will result in a higher number of false positives, erroneous bases called as true variants because of misleading error rate. Other factors are also important such as sequencing depth or coverage to ensure that a high enough number of reads are used in order to call a variant. This can also be influenced by sequencing technology: GC content has an effect on read coverage [146], higher GC content resulting in higher coverage.

In light of these issues with sequencing technology, a simple alignment and mismatch approach is not appropriate to call variants as there is a possibility that we might mistakenly take a sequencing error as a true variant. Therefore a more elaborate approach is required. Several approaches are available, one of the most widely used is the Genome Analysis Toolkit (GATK) developed by the Broad Institute [148]. It is a complex pipeline which has several steps to process the data before calling variants and filtering them once they have been called. A set of clear guidelines [149] is given to the user to make sure that the data are processed correctly and produce a robust set of variants. Notably it recalibrates the base mapping score in order to give a better variant call, resulting in a more robust set of variants. The various steps involved in the GATK pipeline are explained in more detail in the material and methods section 2.10.

The crucial step of this algorithm is calling the variants. This is done using the unified genotype caller, a Bayesian algorithm. The algorithm can call variants on several samples at the same time or sample with multi ploidy (ploidy is the number of chromosome present). The likelihood of the genotype AA, AB and BB are calculated using the following equations [149]:

- $P\{D|GT_i\} = \prod P\{D_{i,j}|GT_i\}$
- $P\{D_i|GT_i = AB\} = (P\{D_{i,j}|A\} + P\{D_{i,j}|B\})/2$
- $P\{D_{i,j}|B\} = \varepsilon_{i,j} * P\{B\, is\, true|D_{i,j,}\, is\, miscalled\},\ D_{i,,j} = B,\ otherwise.$

where $P\{D_{i,j}|GT_i\}$ is the probability of observing $D_{i,,j}$under the hypothesized genotype GT_i. $P\{D_{i,j}|A\}$ and $P\{D_{i,j}|B\}$ are the probabilities of observing base $D_{i,j}$given that the true genotype is either A or B. $\varepsilon_{i,j}$is the probability of miscalling the base of interest and depends on the quality score of the base. Finally $P\{B\, is\, true|D_{i,j,}\, is\, miscalled\}$ is the probability of B being the true chromosomal base given that $D_{i,j}$ is a miscall.

If q_i is defined as $q_i = \{0, 1, 2\}$and is the number of alternative B alleles carried by the single individual i and that $q = \sum_i^N q_i$ is the number of chromosomes carrying the B allele among all the individuals studied or all the members of a pool. Then $P(q = X)$ can be estimated by:

- $P\{q = X|D\} = \frac{P\{q=X\}P\{D|q=X\}}{\sum_Y P\{D|q=Y\}}$
- $P\{q = X\} = 1 - \theta \sum_{i=1}^{2N} 1/i\ X > 0\ otherwise.$
- $P\{D|q = X\} = \sum_{GT\epsilon\Gamma} \prod_i^N P\{D_i|GT\}_i$

- $\Gamma = \{GT\, where\, \sum_i qi = X\}$

With $\Gamma$is the set of all genotype assignments for the N individuals that contain exactly $q = X$ B alleles. $P\{q = X\}$is the infinite-site neutral expectation to observe X alternative alleles in 2N chromosomes with an heterozygosity rate of $\theta$and $GT_i$and D_i are the "ith" individuals' genotype and sequencing read respectively.

Once the variants have been called the next step is to identify differences between our two phenotypes and to look at them in their genomic context in order to identify the potential effect they can have on the data. Several annotation resources and data handling pipelines are used in order to accurately annotate the variants.

1.7 Pig genome release 10.2

The version of the genome used for this project is *Sus scrofa* release 10.2, by Swine Genome Sequencing Consortium (SGSC) [150]. It was released in August 2011 and used a Duroc as the reference pig. A Bacterial Artificial Chromosome (BAC) library was used for the sequencing of the genome. BAC libraries [151] are a collection of bacteria each with an insert of foreign DNA of up to 200,000bp than can be sequenced. BAC are used to construct a physical map of the genome by fingerprinting each clone with an enzyme, allowing the identification of overlapping features and order the BACs. The next step is to sequence the BACs using short reads which will help the assembly as the order of the BACs is already known. The genome was produced from the BAC libraries using short read sequences from the Illumina Genome Analyser II and produced 9906 scaffolds, resulting in 20 assembled chromosomes and 4,562 unplaced scaffolds.

Scaffolds are made of contigs and gaps. Contigs are continuous sequences assembled from short reads and the gaps are usually of know length as they are defined by pair of reads or mate pairs, as shown on figure 1.10. The sequence of the gaps is unknown.

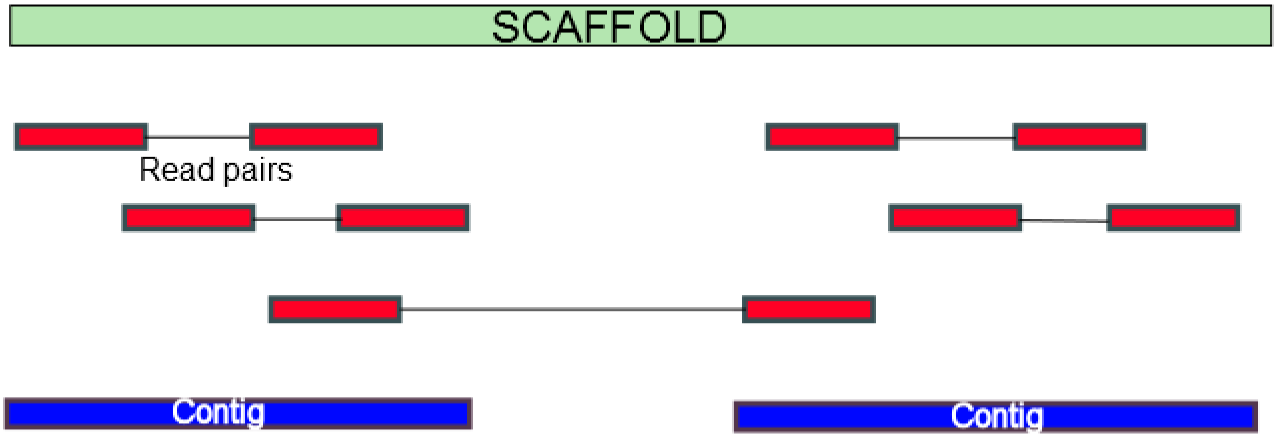


Figure 1.10: Scaffold and contigs. The reads (known sequence in red) form the 2 contigs for this region, the gap is covered by a pair of reads and as the insert size is known a scaffold can be built including the 2 contigs and 1 gap.

The total number of base pairs in the genome is 3,024,658,544 with a golden path size of 2,808,525,991 base pairs. The golden path represents the size of the assembled scaffolds in a continuous sequence, ignoring redundant regions such as pseudo autosomal region and repeat regions (for example in immune genes). Unplaced scaffolds are not counted in the golden path. The scaffolds N50 is 576,008 base pairs and the L50 is 1,303 scaffolds. N50 is the smallest size scaffold producing 50% of the genome in the assembly. In other word 50% of the assembly is comprised of scaffolds at least as big as the N50 value or larger. The L50 is the smallest number of scaffold that can produce the N50 value. The N50 value is small for this assembly, it represents less than 1% (0.2%) of the golden path and the L50 is relatively large, meaning that the assembly is composed of a large number of relatively small scaffolds. If compared to the human genome release at the same time (HG38), the N50 for scaffolds was 59,364,414bp with a L50 of 17 scaffolds. The N50 for the human genome is 1.8% of its total size (3.2 GB).

In terms of the contigs, for the pig a total of 243,021 contigs form the scaffold, the N50 of the contigs is 69,503bp and the L50 is 8,632 contigs. Again the number for the N50 is small and the L50 is large, therefore the scaffolds are formed by a large number of small contigs. In comparisons for the human, the contigs N50 is close to the scaffold N50 at 56,413,054bp and the contigs L50 is 19 contigs.

In light of these numbers, compared to human genome, the pig genome assembly is still very patchy and not to the same standard. The main issue is the size of number of pieces (scaffolds and contigs) making the genome, which is a consequence of the technology used for the sequencing: short reads are more difficult to use for accurate assembly. It is likely that the assembly is not highly accurate, therefore the positioning and annotation of features based on it is not going to be truly representative of the real genome. Furthermore the breed chosen is a Duroc and our data includes Landrace and Large

White which will have a different genetic make up compared to Duroc. Despite these drawbacks, having a reference genome, even if inaccurate, is highly preferable to having to use de novo approaches which take a large amount of time and are more costly to analyse.

The annotation of the genome was performed using the Ensembl automatic gene annotation system [152], which incorporates some of the RNAseq data generated by the SGSC and data from other sources, mainly public databases. The process is divided into three different stages, a raw compute stage, a targeted stage and a similarity stage. The raw compute stage starts with masking repeats using RepeatMasker [153]and Dust [154], in total 48.2% of the genome was masked. After masking several tools are used to identify transcription start sites and then Genescan [155] is used to identify genes structures. To check the accuracy of the prediction, the results of Genescan are aligned against several databases: UniProt [156], Unigene and Vertebrate RNA. For the targeted stage, sequences of known pig proteins are obtained from several sources (UniProt, SwissProt and Genbank [157]) and used to predict coding models using Genewise. Exonerate [158] is used on cDNA, EST and ENSSSCP models to refine the results. The similarity stage uses the protein data from UniProt Protein Existence classification of level 1 and 2, corresponding to protein with experimental evidence at the protein level (1) and at the transcript level (2). Taxonomy is also used to divided the protein models found into different groups: mammalian, non-mammalian vertebrate and invertebrates. Only protein models corresponding to mammalian and non-mammalian vertebrates proteins are kept. Models matching invertebrate proteins are discarded. All of those three steps and some additional evidence generated using pig cDNA, EST (Expressed Sequence Tag) and RNA sequencing data produced by the SGSC were used to finalise the annotation of the genomes. The combined result of this approach defines 21,630 coding genes, 3,124 non coding genes, 568 pseudogenes and a total of 30,585 gene transcripts.

1.8 Objectives and aim

The aim of this study is to get an understanding and identify the genetic factors that contribute to maternal infanticide in pigs. Previous work done on this trait has established that there is a genetic and heritable component to it. The work done by Quilter et al ,[17, 80, 2] and Chen et al [14] have highlighted genes and regions of the genome that might influence maternal infanticide in pigs. These studies used a range of methods: linkage using microsatellite, gene expression and genome wide association study using microarrays. The evolution in the technology available allows us to further complete some of the work done by using more advanced methods and increase the resolution at which we look at the regions they identified as shown in figure 1.11. Using some of the previous work as a starting point, notably Quilter et al [2], this study will use new data and combine it with existing data from Quilter et al [2] to identify regions of interest using more powerful methods to investigate families. Once regions of interest are defined, sequence capture will be used to sequence and call sequence variants in these regions. For the sequencing, animals will need to be selected in order to constitute pool of infanticide and control animals. The variant allele frequency between the two groups will be compared in order to identify variants of interest. Once identified the variants will be annotated and their potential impact on the region or genes assessed. An overview of the study is shown on figure 1.12.

The following objectives can be defined from the works and methods covered in the introduction:

- Add new animals to the previous data generated by Quilter et al [2] in order to generate pedigree to use parent of origin and family base association test
- Identify new regions of interest from the family based association and parent of origin test
- Design capture sets to target the regions of interest
- Select infanticide and control animals for sequencing
- Sequence the regions of interest
- Process the sequenced data to call variants
- Compare allele frequencies between infanticide and control animals to identify variants of interest
- Use resources from pig and human databases to get the most accurate annotation for the variants of interest
- Analyse the variant to identify specific genes or precise locations of the genome linked to maternal infanticide
- Assess the potential impact of variants in genes of interest

Combining different approaches of genotyping and sequencing will help narrow down the regions and genes involved in the genetic make up of maternal infanticide. Methods used in the previous study such as microsatellite markers have helped identify regions of the genome but with the advances in methodology such as microarray and sequencing, it is possible to investigate in greater details the genome and identified regions a few kilobases large instead of larger scale in centimorgan. The access to better methods to analyse the data set available will also help increase the power of the analysis to pick good candidate regions for variant analysis. If successful this study will identify key genes and causal variants that are significantly contributing to maternal infanticide in pigs.

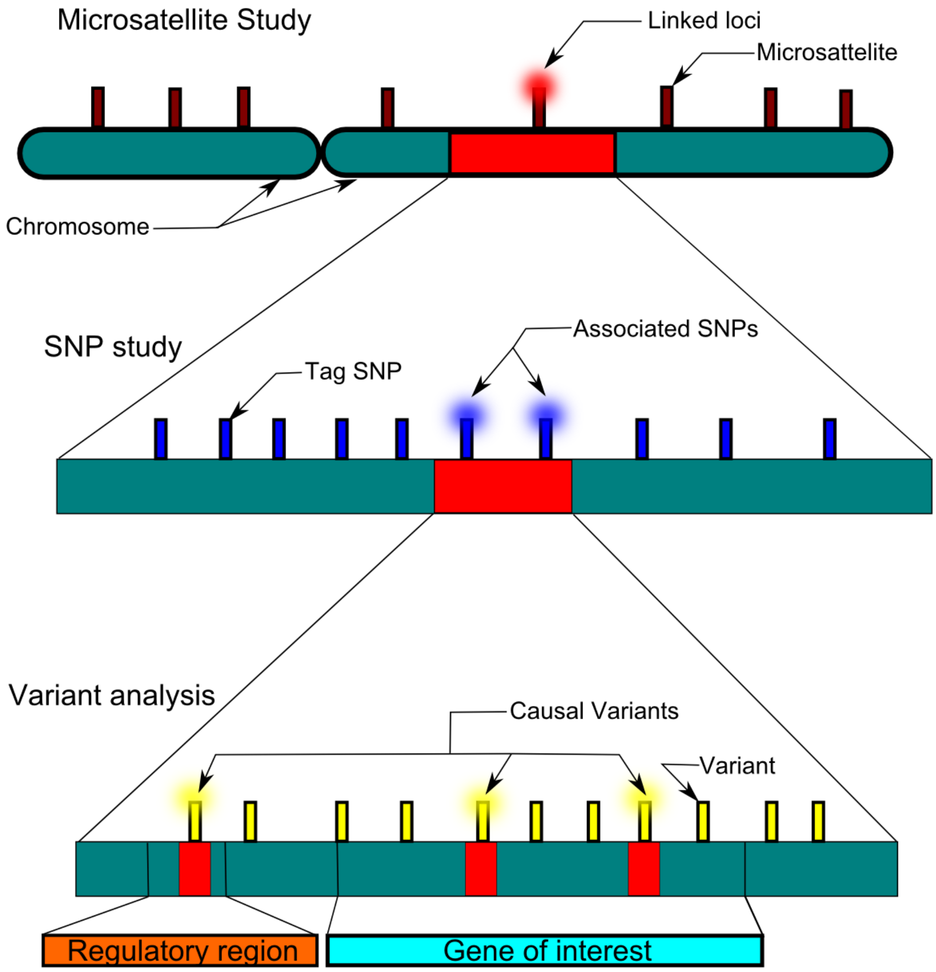


Figure 1.11: Overview of the different methods to identify causal loci on the genome. The resolution of the methods improve greatly from microsatellite (scale in centimorgans), to association study (scale in mega bases) to variant analysis (scale in kilobases or less).

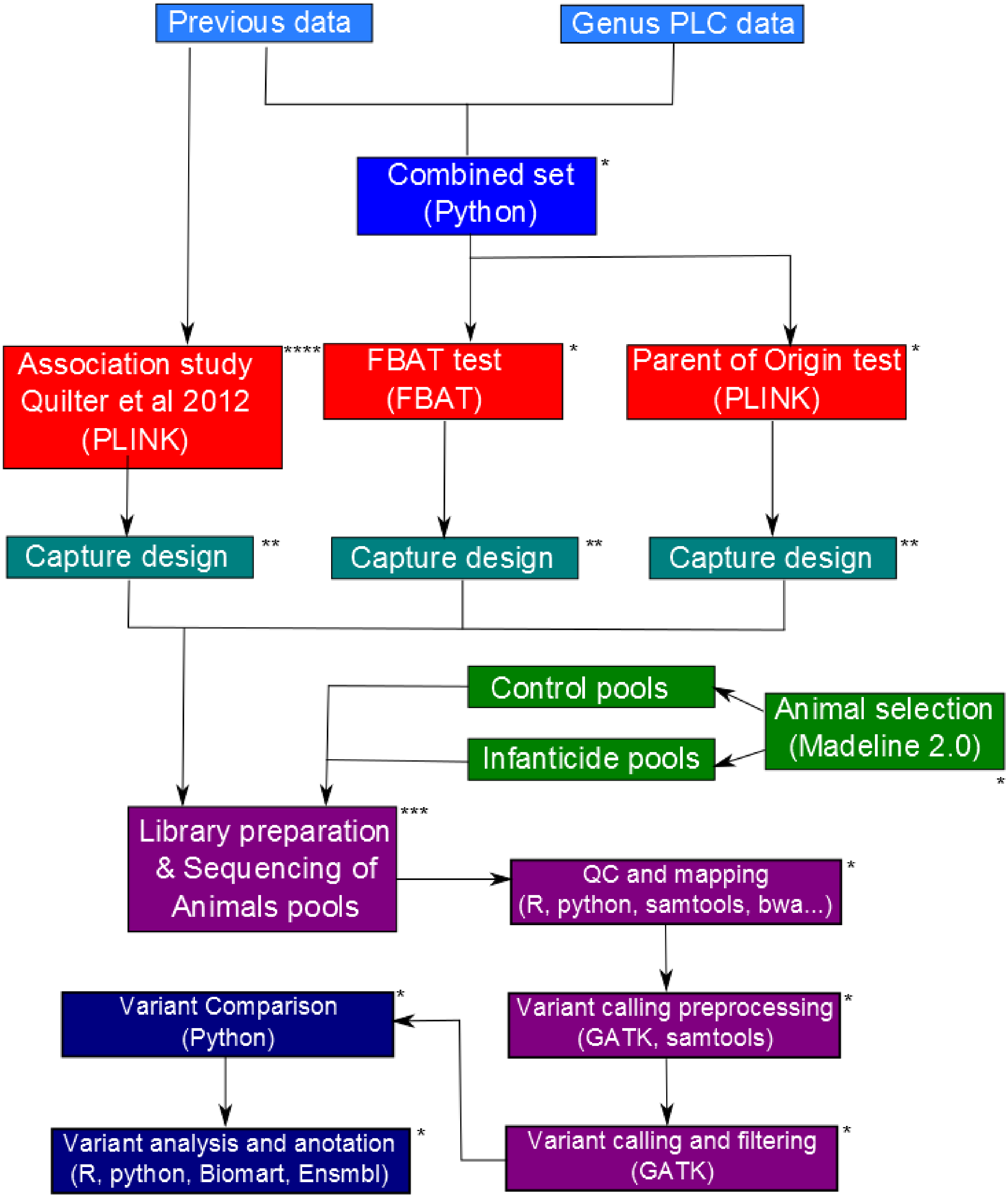

Previous data
Genus PLC data
Combined set (Python) *
Association study Quilter et al 2012 (PLINK) ****
FBAT test (FBAT) *
Parent of Origin test (PLINK) *
Capture design **
Capture design **
Capture design **
Control pools
Infanticide pools
Animal selection (Madeline 2.0) *
Library preparation & Sequencing of Animals pools ***
QC and mapping (R, python, samtools, bwa...) *
Variant Comparison (Python) *
Variant calling preprocessing (GATK, samtools) *
Variant analysis and anotation (R, python, Biomart, Ensmbl) *
Variant calling and filtering (GATK) *

Figure 1.12: Overview of the study. The number of "*" highlight the various contributions: * Julien Bauer, ** Claire Quilter, *** Kerry Harvey, **** Multiple contributions, see [2].

2 Material and Methods

2.1 Animals for the family based association and parent of origin tests

All DNA samples used in this study were extracted from animals selected from herds maintained and catalogued by Genus PLC. Detailed pedigree information was provided by Genus where available. The different lines are as follow: line B is a Landrace line, C is a Large White line, D is a cross between Landrace and Duroc and H is a cross between Large White and Duroc. Line A and B are dam lines, selected for their good maternal instincts and litter sizes. Line D and H are both sire lines, selected for their meat quality traits. The incidence of infanticide events for these lines is 4.8% for line B, 5.9% for line C, 10.8% for line D and 10.3% for line H. In total 1429 animals were typed in Cambridge, 309 for line B, 219 for line C, 423 for line D and 478 for line H, more information is available in table 3.1. Genus provided data for a further 208 animals for line B, 189 for line C and 150 for line H, more details about the animals are available in table 3.2. The number of families available for the family based association is 76 for line B, 50 for line C, 56 for line D, 96 for line H. Details about the animals in those families is given in table 3.5.

2.2 Summary of bioinformatics tools used for the analysis

2.2.1 Genotyping

- PLINK [159] was used for the quality control of the genotyping data and running the parent of origin analysis (see below)
- FBAT [50] was used to perform the family based association test
- GenomeStudio (Illumina proprietary software) was used to load the data generated from the array and generate PLINK formatted output for the analysis
- Python scripts were used to merge the data provided by Genus to the data generated and already available in Cambridge.

2.2.2 Sequencing data processing

- FastQC [160] was used to generate quality metrics from the sequencing run.

- The R [161] software and more specifically the TEQC [162] and ggplot2 [163] packages were used to generate graphs and quality metrics for the sequencing data and assess the efficiency of the capture sets.

- BWA [144] was used to align the data to the reference genome obtained from Ensembl [152], the version of the pig genome used was version 10.2 at the time of the work. “BWA mem” was used, it is the most recent implementation of the BWA algorithm.

- Samtools [164] was used to manipulate the alignment files (bam) after the alignment in order to sort the file by chromosomal order, generate index file for further processing or for general investigation of any issues or queries related to them.

- PicardTools [165]was used for the specific preparation of the data for processing in the Genome Analysis Tool Kit (GATK), this includes generating a dictionary for the reference file, marking duplicates reads, adding annotation to the alignment files. It was also used to extract some quality metrics from the aligned files.

- GATK [148] was used to perform the variant calling. Several steps are involved in the process and are described in more detail in section 2.10. This tool is often used in conjunction with PicardTools.

- Python scripts were used to compared the variant calling files from the different pools.

2.2.3 Annotation of the comparisons

- NCBI e-utilities, specifically the Python API library, was used to retrieve the six hundred base pairs of sequence around each pig variant identified by the comparisons (300 base pairs either side of the SNP), see section 2.11 for more details on the comparisons.

- Blast [166]was used to increase the accuracy of the synteny mapping with the human using the sequences retrieved. More details about this in section 2.11.3.

- Blastp [167] was used to compare the protein sequence of genes around amino acid substitution.

- The NCBI Python API was also used to parse the blast results and extract the positions of the matches (of the blast) against the human

- R and the biomaRt [168, 169] module was used to retrieve annotations from the data, using the SNP databases for human and pig, and updating data for the pig genome from SS 10.2 to SS 11.

- The Ensembl Perl compara API was also used to get syntenic regions: using the pig coordinates to retrieve the corresponding syntenic blocks in the human.

2.2.4 Data manipulation

- Python [170] scripts were used for data manipulation such as merging files, extracting data, adding data and sorting files.
- The R package dplyr [171] was used to merge data in R, more specifically for merging genomic intervals or tables with a large number of entries.

2.2.5 Genomic and pedigree plotting

- The R package Gviz [172] was used to plot genomic regions for the pig and human.
- The pedigree software used to draw the pedigree in order to select the animals is Madeline 2.0 [173]
- ggplot2 [174] was used to generate some of the QC plots
- TEQC [162] was used to generate plots for the sequencing capture QC
- QQman [175] was used to plot the Manhattan and QQ plots for the results of the FBAT study

2.3 Generation of the microarray data

This part of the work was performed by Claire Quilter and Kerry Harvey.

Each DNA sample was extracted and processed according to the manufacturer's instructions (Illumina Inc, San Diego CA) for hybridisation on the microarray. The set of samples processed at Cambridge were hybridised on the Porcine SNP60k version1 array by Illumina. The later set of data generated at Cambridge was hybridised on the SNP60K version 2 by Illumina. This updated version of the array contains a pool of common probes with the previous version of the array. It is possible to combine both types by reducing the set analysed to the part common between both arrays.

The processing of both arrays versions followed a similar protocol: the extracted DNA is whole genome amplified, then fragmented and hybridised overnight to the array. The hybridised product goes through a one base extension for each probe to type the SNP of interest. Each SNP will be represented by two different probes on the array, one for each allele. Once the extension step is completed the arrays are stained using two fluorochromes, then scanned on the iScan Illumina scanner. The scanned data are loaded into GenomeStudio for technical quality control. The data were then exported and version 1 and version 2 data were merged together, resulting in a set of 1522 samples: of these 54 were duplicate animals. In addition to this, the data provided by Genus added 547 animals to the study for a total of 1976 animals. The Genus set was merged with the Cambridge set to provide the data set used for the FBAT and PO tests.

2.4 PLINK data format

The file format used for the PO and FBAT analysis, see section 2.6 and 2.7, is the PLINK format. The PLINK format for the analysis consists in two files, the PED file and MAP file. The PED file consists of the meta data columns followed by the SNP data. The meta data columns in the file are:

1. Family identifier
2. Individual identifier
3. Paternal identifier
4. Maternal identifier
5. Sex (1 = male, 2 = female, other = unknown)
6. Phenotype of interest

After the meta data columns, each genotype has two columns for each SNP, one for each allele present in the individual.

The MAP file contains the coordinate of the SNP present in the PED file, the order of the rows in the MAP file should match the order of the SNP in the PED file.

The MAP file columns for the SNP coordinates are:

1. Chromosome
2. SNP identifier
3. Genetics distances (morgans)
4. Base-pair position (bp units)

For each of the lines studied (see section 2.1), a specific PED file was generated in order to test each line apart from the other. The MAP file was common for all of the lines. Note that the FBAT software need the MAP file to start with the SNP identifier.

2.5 Data quality control

The quality control for the merged data sets was performed in PLINK [159], the arguments used were: *geno* 0.05, *hwe* 0.0001, *mind* 0.1. The first argument set the genotype missing rate, it is used to remove any SNPs which were not called in 5% or more of the samples. The second is the threshold for the Hardy Weinberg equilibrium test. It is used to test the distribution of the markers in the population as it should follow the equation: $p^2 + 2pq + q^2 = 1$ with p the frequency of the A allele, q the frequency of the B allele for a AB genotype. Any marker deviating from this might be affected by population stratification or admixture and should be filtered out. The third parameter is the missing rate for individual samples. If a given sample has more than 10% of the SNPs not called it will be excluded from the analysis. This QC was run for each line on the combined set of data set, combining the data generated in Cambridge and provided by Genus PLC.

2.6 Parent of Origin analysis

The parent of origin analysis was performed in PLINK [159] using the filtered data set as described above, see section 2.3, using the following options in PLINK: *poo,* and *tdt* in order to use the transmission disequilibrium test. The parent of origin analysis is similar to a normal transmission disequilibrium test but considers the transmission from a heterozygous mother and father separately in order to identify preferential transmission from one of the parents.

2.7 Family Based Association study

The family based association study was performed using the software FBAT based on the work in [50]. The input data is in the PLINK format, see section 2.4. Once the data loaded into the software, the comparisons were performed using the 'fbat' command, setting the minimum number of informative families to 10. The results were saved to a text file that was used for interpretation.

2.8 Selection of region of interest

The selection of the regions of interest and the design of the three sequence capture panels was done by Claire Quilter based on the results from the previous study [2] and results from the work presented here, the Family Based Association Test (see sections 2.7 and 3.1.5) and the Parent of Origin test (see sections 2.6 and 3.1.6).

From the previous study [2], the top haplotype regions, with a $-log10(p_{value})$ above 4, were selected to design a capture set, called Capture 1. The regions selected are on chromosome 1 (65MB and 74MB), chromosome 3 (26-27MB, 29-30MB, 36-37MB, 100MB and 131MB), chromosome 4 (1.9-2.1MB, 5-5.2MB and 121MB), chromosome 6 (21MB), chromosome 10 (10MB), chromosome 12 (34MB, 62MB), chromosome 13 (22-23MB), chromosome 14 (24MB), chromosome 15 (17MB, 130MB, 137MB), chromosome 18 (11MB and 18MB) . One of the regions for chromosome 3 was extended to included the GP2/GPR139 region as it is a candidate region for puerperal psychosis. Figure 2.1 shows the number of probes per chromosome, figure 2.2a and 2.2b the size of the probes across each chromosome. In total the combined size of the targetted regions is 7,466,508 bases using *Sus scrofa* genome version 10.2 and 7,244,762 bases for *Sus scrofa* genome version 11.1 (via coordinates lift-over).

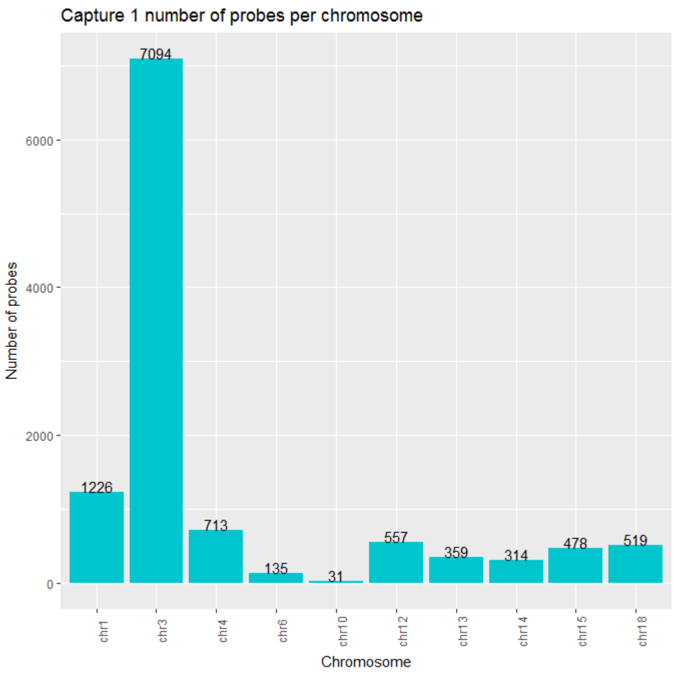
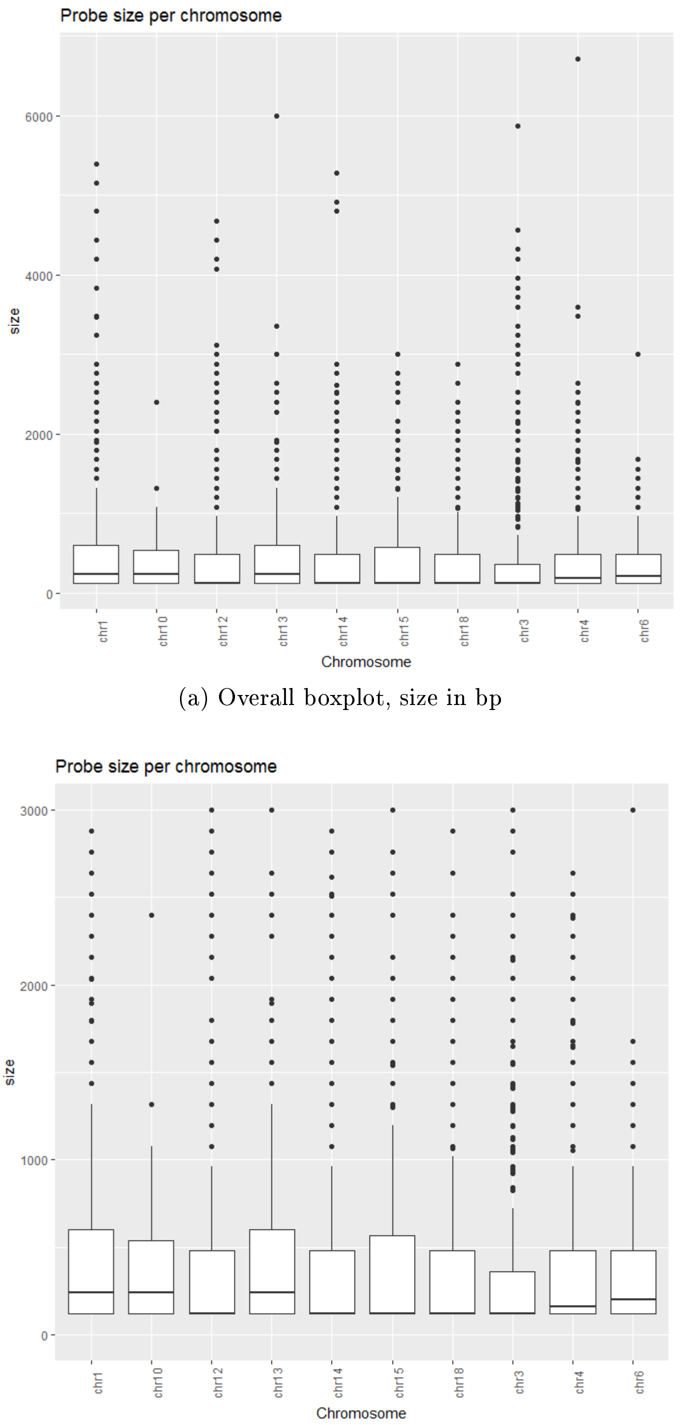


Figure 2.1: Capture set 1: number of probes per chromosome. A probe is defined as a sequencing interval being captured by the panel



(a) Overall boxplot, size in bp

(b) Zoomed boxplot, size in bp

Figure 2.2: Capture set 1: boxplots of the probes size per chromosome. 2.2a: overall boxplot, 2.2b: zoomed, excluding probes bigger than 3000 base pairs.

The second capture set, Capture 2, was selected using the results from the family based association test, it includes all the regions that returned a $-log10(p_{value})$ above 4. It also includes regions that were identified in a RNA sequencing pilot experiment [176] from on going work in the group. The genes selected were the top 30 genes in the RNA sequencing comparison. The regions selected are on chromosome 1 (13MB, 106MB, 121MB. 129MB, 135MB, 150MB, 191-192MB, 228MB, 244MB, 247MB and 294-295MB), chromosome 2 (27MB, 36MB, 42MB, 82MB), chromosome 3 (22MB, 45MB, 75MB), chromosome 4 (4MB, 16MB, 83MB, 97MB, 113MB), chromosome 5 (3MB), chromosome 6 (21MB, 83MB, 90MB, 108MB, 112MB, 123MB, 157MB), chromosome 7 (3MB, 7MB, 9MB, 18MB, 23MB, 34MB, 59MB, 79-80MB, 89MB, 95MB, 106MB, 132MB), chromosome 8 (118-120MB), chromosome 9 (5MB, 9MB), chromosome 10 (16MB, 24MB, 43MB, 50MB), chromosome 11 (8MB, 21MB, 24MB), chromosome 12 (10MB, 23MB, 54MB), chromosome 13 (11MB, 22MB, 45MB, 60MB, 149Mb, 190MB, 194MB, 202MB, 208MB, 213MB, 215MB), chromosome 14 (3MB, 15MB, 52MB, 66MB, 70MB, 126MB, 130MB, 134MB, 138MB, 150-151MB), chromosome 15 (11MB, 12MB, 22MB, 45MB, 132MB, 147MB), chromosome 16 (21MB, 41MB, 44MB, 47MB, 53MB, 55MB, 85MB), chromosome 17 (37MB), chromosome 18 (14MB, 26MB, 41MB, 43MB, 50MB), chromosome X (113MB). Figure 2.3 show the number of probes per chromosome; figure 2.4a and 2.4b the size of the probes across each chromosome. In total the combined size of the targetted regions is 10,351,334 bases using *Sus scrofa* genome version 10.2 and 9,792,664 bases for *Sus scrofa* genome version 11.1 (via coordinates lift-over).

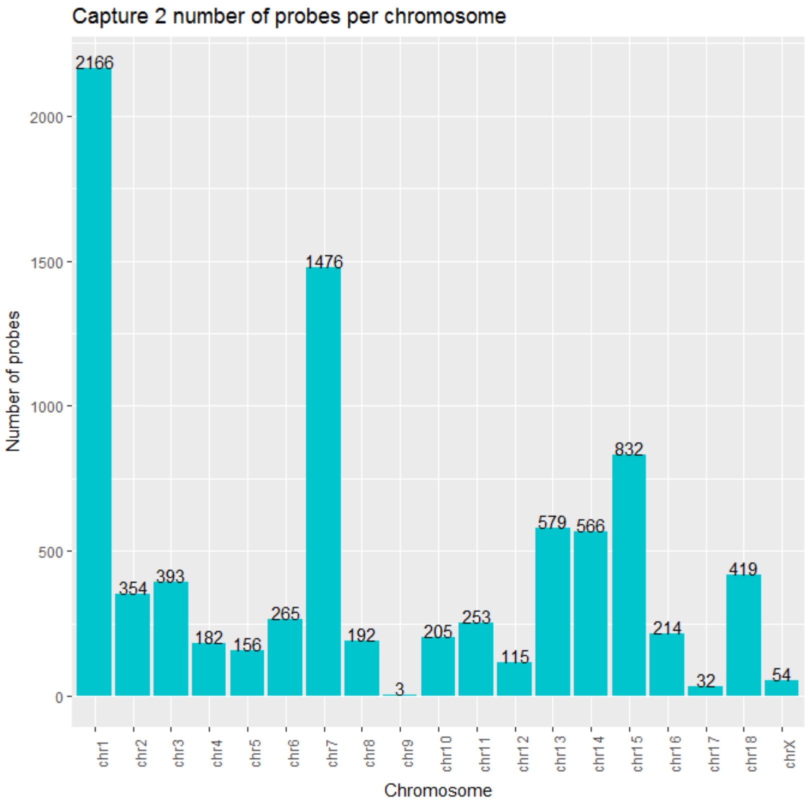
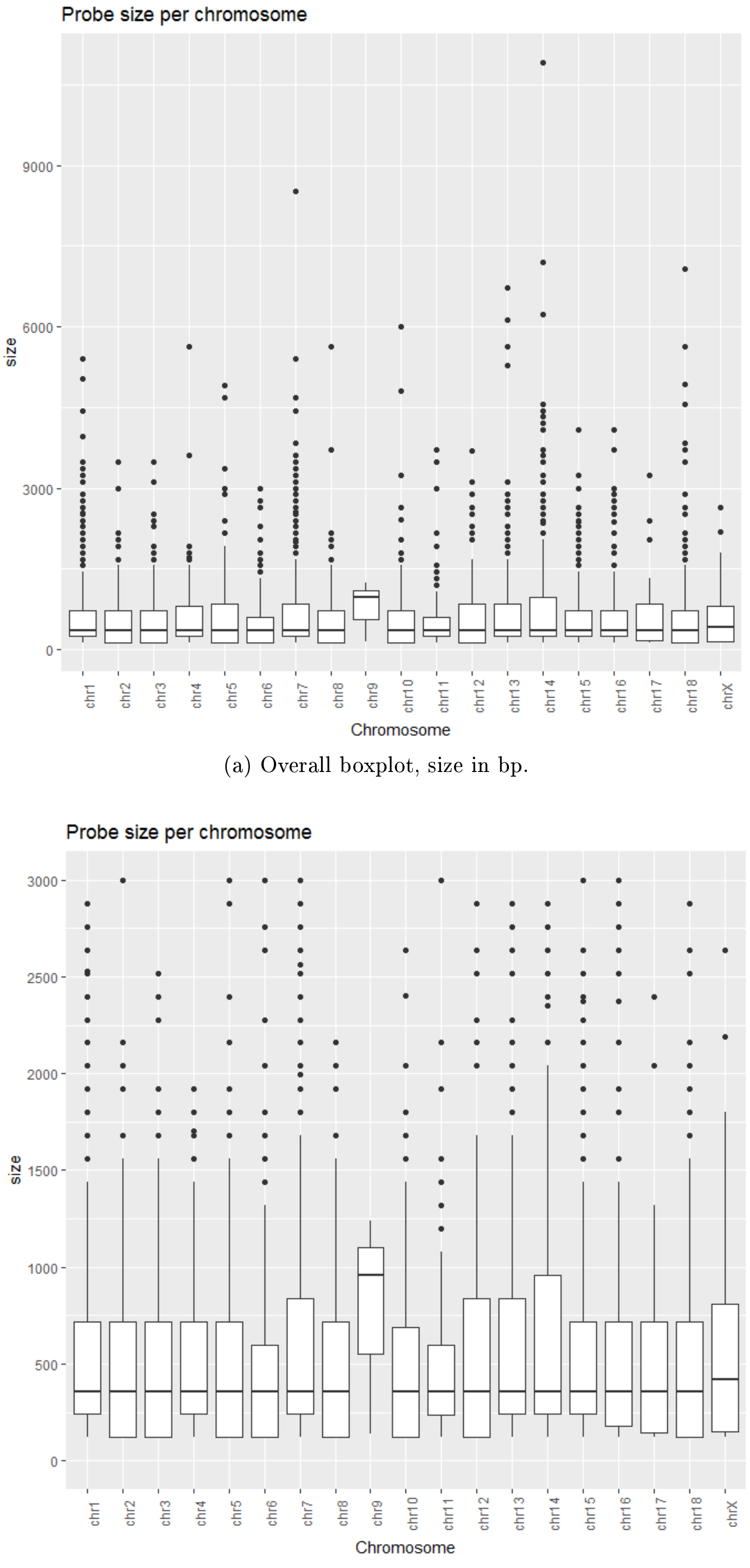


Figure 2.3: Capture set 2: number of probes per chromosome. A probe is defined as a sequencing interval being captured by the panel



(a) Overall boxplot, size in bp.

(b) Zoomed boxplot size in bp

Figure 2.4: Boxplots of the probe size per chromosome for capture set 2, 2.4a: overall boxplot, 2.4b: excluding probes bigger than 3000 base pairs.

The third and final capture set, Capture 3, was designed from the results of the Parent of Origin

(PO) results. The regions were selected by filtering the results using a maternal chi square p-value below 0.003, with the exception of 4 gene regions, 3 from chromosome 3 and one from chromosome 2. These region were over the selection threshold but not significantly and contained gene of interest. The regions captured are, chromosome 1 (29MB, 45MB, 65MB, 76MB, 92MB, 96MB, 102MB, 139MB, 181MB, 191MB, 193MB, 200MB, 250MB, 284MB), chromosome 2 (6MB, 12MB, 14MB, 27MB, 86MB, 112MB, 121MB, 125MB, 138MB), chromosome 3 (7MB, 14MB, 17-18MB, 19MB, 30MB, 31MB, 32MB, 45MB, 75MB, 106MB, 137MB), chromosome 4 (2MB, 10MB, 21MB, 33MB, 38MB, 77MB, 80MB, 83MB, 105MB, 119-120MB), chromosome 5 (1MB, 7MB, 66MB, 67MB, 99MB), chromosome 6 (14MB, 23MB, 38MB, 112MB, 157MB), chromosome 7 (63MB, 75MB, 95MB), chromosome 8 (15-16MB, 89MB, 94MB, 117MB, 138MB), chromosome 9 (26MB, 33MB, 73MB, 102MB, 116MB, 120MB, 121MB, 128MB, 139-140MB, 149MB) chromosome 10 (36MB, 52MB, 55MB), chromosome 11 (22MB, 66MB), chromosome 12 (16MB, 41MB, 52MB), chromosome 13 (1MB, 7MB, 26MB, 38MB, 76MB, 77MB, 189-190MB, 202MB), chromosome 14 (17MB, 18MB, 34-35MB, 62MB, 79MB, 101MB, 113MB, 129MB), chromosome 15 (19MB, 46MB, 52MB, 143MB), chromosome 16 (32MB, 72MB, 74-75MB), chromosome 17 (19MB, 23MB, 46MB, 53MB), chromosome 18 (1MB, 2MB, 8MB, 24MB, 49-50MB). Figure 2.3 shows the number of probes per chromosome, figure 2.4a and 2.4b the size of the probes across each chromosome. In total the combined size of the targetted regions is 13,347,528 bases using *Sus scrofa* genome version 10.2 and 13,876,030 bases for *Sus scrofa* genome version 11.1 (via coordinates lift-over).

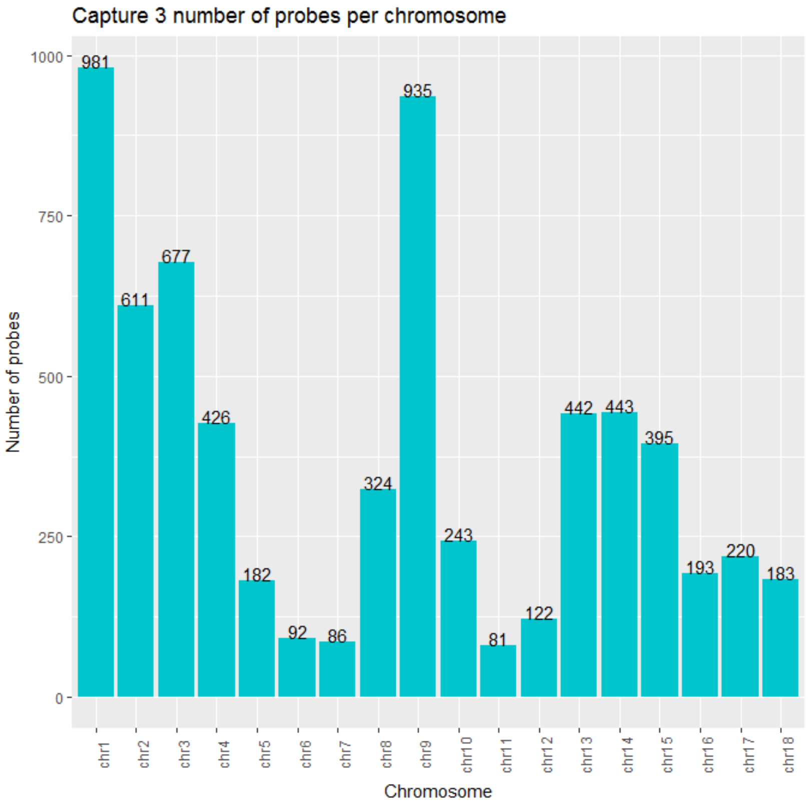
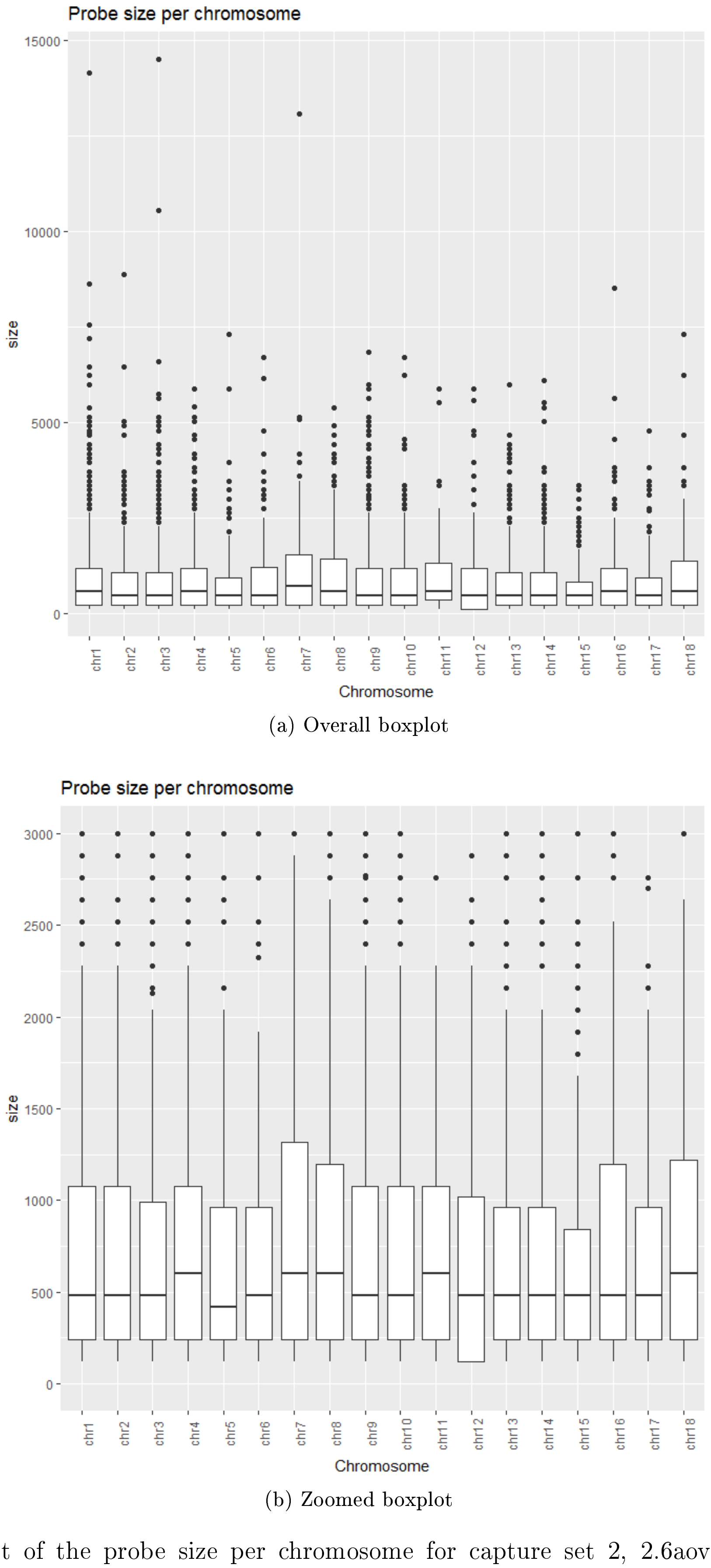


Figure 2.5: Capture set 3: number of probes per chromosome. A probe is defined as an sequencing interval being captured by the pane



(a) Overall boxplot

(b) Zoomed boxplot

Figure 2.6: Boxplot of the probe size per chromosome for capture set 2, 2.6aoverall boxplot, 2.6b boxplot removing probe longer than 3000 base pairs

2.9 Sequencing of the regions of interest

2.9.1 Selection of the animals

Selection of the animals for sequencing was done using the pedigree information available. The animals were segregated in individual families, the family tree was drawn using the software Madeline 2.0 [173].

In addition to having had at least one instance of infanticide, two different criteria were applied to select individuals:

1. Having an history of infanticide in the family. At least one of the female ancestors or descendants of the selected animals need to have an incidence of infanticide.

2. Having several instances of infanticide. At least two episodes of infanticide are necessary to be considered in this category.

For the first criterion the animals were selected by generating pedigree trees for each of the families in the study. An example of a relatively simple pedigree is shown in figure 2.7 while figure 2.8 shows a more complex one. The rest of the pedigrees used are available in the supplementary figures.

In the example on figure 2.7, both selected animals have an history of aggression, they are cousins and one of their dams (52832350) was also aggressive.

For the second criteria, we used the phenotypic information available to select animals which had killed piglets in at least two litters.

Unfortunately not all the selected animals had enough DNA left to be used for sequence capture.

The number of animals available to build the pools for each line is summarized in table 2.1. For line B and C the pools are as follow: pool one is composed of animals with an history of infanticide in their family, pool two of animals that are serial offenders and pools three and four are the controls. For line D and H the animals with an history of infanticide were run in three different pools. For line D pool one, two and three were comprised of 24, 16 and 16 animals, pool four is for serial offenders and five and six the controls. For line H the first 3 pools were also comprised of individuals with an history of infanticide, pool one, two and three comprised 10, 13 and 9 individuals, pool four is the serial offender pool, and five and six the controls. The animals have to be divided into pools due to the technical limits of the sequencer and in order to achieve sufficient coverage.

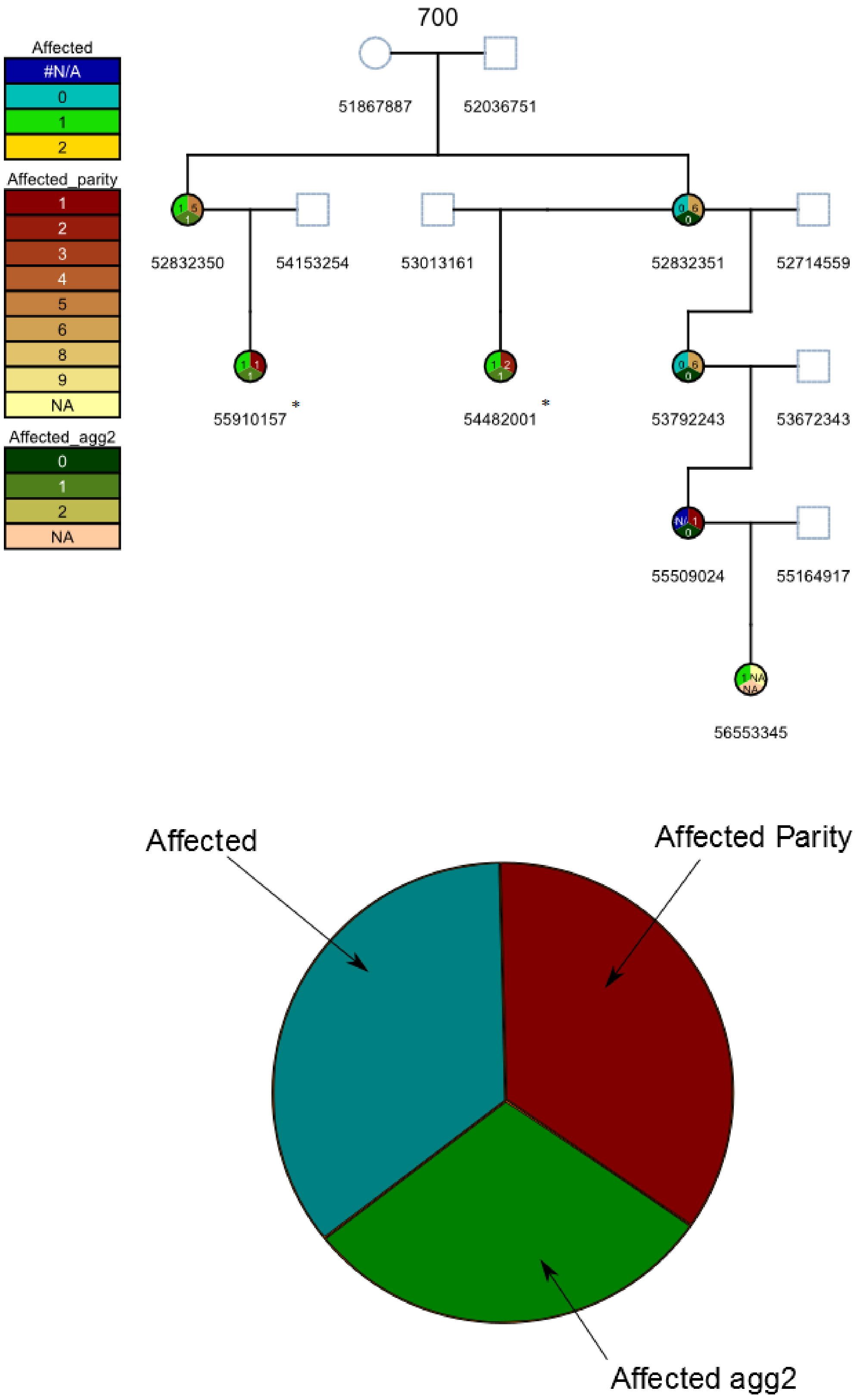


Figure 2.7: Example of pedigree generated using Madeline 2.0. Family 700 for line B. The two selected samples are denoted by a *. Each individual has three possible different sources of phenotypic data, represented by a pie chart. A large version of the pie chart is displayed below the pedigree for clarity. The "Affected" number comes the first data set phenotypic data and is sometime incomplete, 0 is for non infanticide, 1 for infanticide. The "affected agg2" number follows the same convention and comes from the second set of annotations, that was generally more complete. The "Affected_parity" gives us the number of parities the sow had, regardless of any infanticide event. The main difference between those phenotypic data is that the information about the history of the samples typed was sometimes not provided or accurate in the first set of phenotypic data.

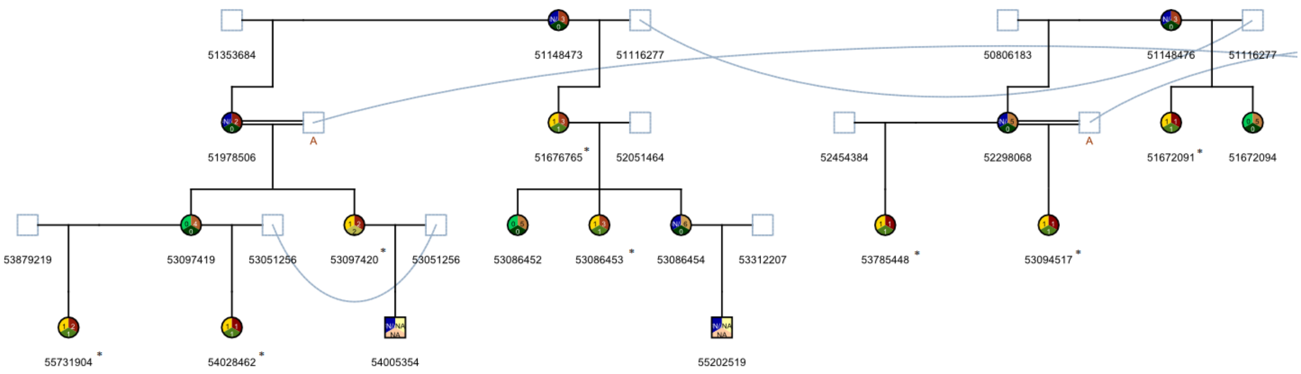


Figure 2.8: Example of a more complex pedigree/family. For line H, this pedigree displays an example of sharing of sires for this line.

	Control pool 1	Control pool 2	Serial infanticide	History of infanticide in the pedigree
Line B	25	25	6	4
Line C	25	25	4	12
Line D	22	22	14	46
Line H	25	25	4	32

Table 2.1: Animals pools

2.9.2 Library preparation and sequencing

The library preparation was done by Kerry Harvey and the sequencing done by the team of Cambridge Genomic Services at the Department of Pathology.

The libraries were prepared according to the Agilent Sure Select protocol. Briefly the DNA is sheared and sequencing adapters are added to the DNA fragments, then the fragments are hybridised to DNA baits which are biotinylated. The hybridised DNA is then purified using streptavidin coated magnetic beads. Once the fragments are isolated the DNA baits are digested and the remaining DNA fragments are amplified and ready to be sequenced. The libraries are sequenced using a paired end approach, sequencing a 100 base pairs (bp) section from each end, resulting in two 100 bp reads per fragment. Two raw sequence files are generated for each sample corresponding to read one and read two. For each of the three capture sets, the pools were sequenced in two runs on the Illumina NextSeq® sequencer. A total of 8 sequencing runs were performed. Once the sequencing was completed, the reads generated were passed on for processing and variant calling.

2.10 Read processing and variant analysis using the Genome Analysis Tool Kit (GATK)

Once the sequencing run completed, the read data were converted from the Illumina 'bcl' property format to FastQ format using the bcl2fastq software (proprietary software from Illumina). The reads were processed via the pipeline described in figure 2.9. Starting with raw reads, the data were first put through the preprocessing and mapping step. This was followed by post processing to format the data for variant calling in GATK [148]. The mapping was performed using the BWA mapping software [144] and various checks were done on the data pre and post mapping. Post processing was done using Picard tools [165] and GATK. Once the data were correctly formatted the variants could be called using the Unified Genotype caller of the GATK suite. After calling, the variants were filtered using a set of filters in order to remove poor quality variants. The variant frequencies from the different pools were compared in order to identify markers that are behaving differently between infanticide and normal pools of individuals.

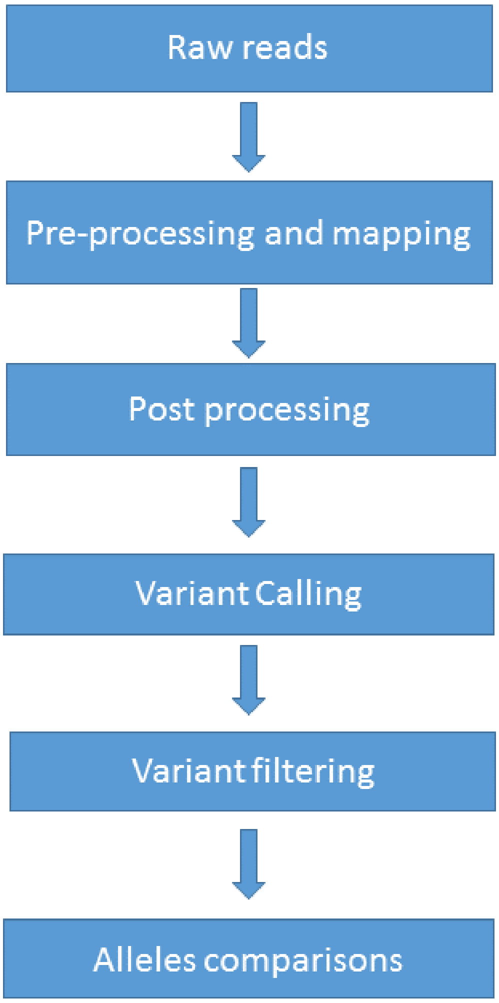


Figure 2.9: Pipeline for sequencing data processing.

2.10.1 Read preprocessing and mapping

The read pre-processing consists of checking the quality of the reads prior to mapping. This is done using FastQC, which generates various graphs and tables in order to assess the quality of the reads.

A usual step to most sequencing data analysis is the trimming of the reads but for variant calling this step is not beneficial (see section 4.2.2 of the discussion), therefore it was skipped. After the raw read quality control, the reads are mapped against the reference genome using the Burrows-Wheeler Aligner (BWA) [144]. The reference genome used for this thesis was *Sus scrofa* 10.2, obtained from the Ensembl website. The genome was download in FASTA format and indexed for use with BWA ("*bwa index*"). The BWA algorithm used for the mapping is "*bwa-mem*". The output format after mapping was Sequence Alignment Map, or SAM, format and was immediately converted to Binary Alignment Map, or BAM, a compressed version of SAM, in order to save space. The SAM format is a tab delimited file consisting of an optional header section and of an alignment section. Header lines are defined by a "@" at the beginning of the line. The alignment lines consists of 11 mandatory columns:

1. QNAME: the query template name
2. FLAG: the bitwise flag, defining the mapping status of the segment
3. RNAME: the reference sequence name
4. POS: the 1 based leftmost mapping position
5. MAPQ: the mapping quality score
6. CIGAR: the CIGAR string, defining the alignment status for each base in the segment
7. RNEXT: reference name of the mate read
8. PNEXT: Position of the mate read
9. TLEN: observed template length
10. SEQ: segment sequence
11. QUAL: Phred-scaled base quality score

Once mapping is completed, post mapping quality control steps are performed in order to determine how well the sequence capture performed and how well the regions captured are covered. The R package Target Enrichment Quality Control (TEQC) [162] was used to generate graphs to evaluate the efficiency of the three capture sets and the quality of the sequencing coverage.

2.10.2 Post mapping processing

Once the mapping steps complete the mapped reads were processed in order to call variants with the GATK suite. The first step in this process is to sort and index the bam files for further processing, this step is performed using samtools. The next step is to mark duplicated reads using PicardTools. Marking duplicate reads is important in order to get a good estimate of allele frequencies for the variants. Duplicated reads are defined as two pairs of reads that have the same start point for both read one and read two. In other words; they represent the same fragment sequenced two times or more, which will bias the variant calling and allele frequency estimation. The next step is to add read group information. This step is required by GATK but is not essential. It consists on adding information to the read files, to include the sample id, instrument used, library and pool. This step is critical for consortium or large studies as the final data set might come from several core labs. An important step performed next is the read realignment. This step is performed in order to correct the alignment of reads around indels and is essential when using the UnifiedGenotype caller for variant calling. It will identify intervals of the genome where indels are present (from a reference file) and realign those regions with a local realignment algorithm that will correct any misalignments due to the presence of indels in the read. This step was performed using IndelRealigner and the reference database used for the indels was *Sus scrofa* dbSNP 145.

Once the reads have been realigned the next step is to perform the base quality score recalibration, or BSQR, using BaseRecalibrator. This process is necessary because variant calling relies on the quality scores attributed to the base by the sequencing instrument. However the instrument does not always attribute the right scores to a base call due to the various sources of technical bias present in the algorithms they use. This recalibration is important as the variant calling algorithm relies heavily on the quality score to identify potential variants. BSQR uses an empirical approach and machine learning in order to adjust the base quality score, which improves the quality of the variant called. This is done in two steps: first the model of covariation is built based on the data and known variants if available, *Sus scrofa* dbSNP 145 was used as the reference, then the base quality scores are adjusted according to the model generated.

Note that for all the post mapping processing and the variant calling, the parameter “-L” for the GATK tool suite was used. This parameter takes as argument a bed file, a file containing the coordinate of target regions in tab delimited format. For each capture set the file used for this argument is the capture set bed file, which contains the coordinates of the probes used for the capture.

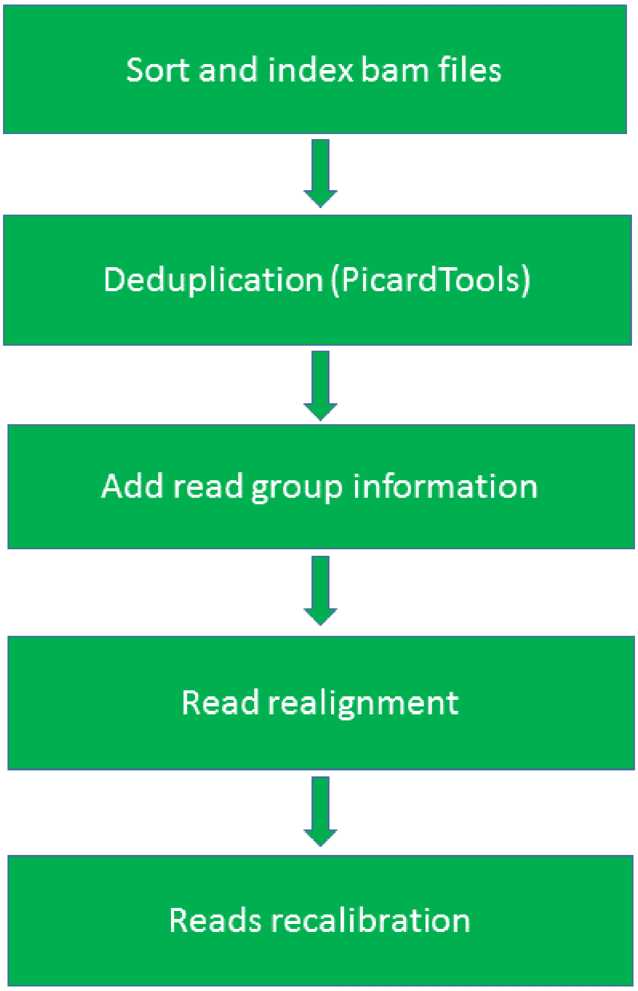


Figure 2.10: Post mapping processing

2.10.3 Variant calling and filtering

The final steps of the GATK pipeline are to perform the variant calling, filter the variants and generate lists. The variant calling was done with the UnifiedGenotype caller, the reasons for this choice are discussed in section 4.3.2.1. This algorithm calls the variant for each sample and can use a reference file to help guide the variant calling. For this thesis the variant reference was the pig SNP database 145. The ploidy of each of the pools was provided and the additional options used were: *stand_call_conf* of 50.0 and *stand_emit_conf* of 10. The first parameter is the minimum phred-scaller confidence threshold for the variant calling which will flag any SNP below that phred score as "filtered". Any variant with a phred score lower than stand_emit_conf of 10 will not be considered and is excluded from the output.

After the variants are called the next step is to remove low quality and potential false positives from the set. The filtering was done using the *VariantFiltration* command and the following parameters:

- QD below 2; Quality by Depth, which is the variant confidence score divided by the unfiltered depth of the non homozygote reference samples, using variant matching the alternate allele. It is used to normalised the variant quality in order to avoid bias due to deep coverage. Any variant with a score below 2 is deemed to have a score too low compared to the depth of coverage based on empirical evidence.

- FS above 60. FS represents the Fisher Strand which is used to asses the potential strand bias in the variant. FS represent the phred-scaled probability of having strand bias at the variant locus, that is the probability of observing a variant more or less often on the forward or reverse strand. A locus with no strand bias will have a FS value close to 0. However most loci tend to have some strand bias and based on empirical evidence the default value is set at 60.

- MQ below 40; MQ is the root mean square of the mapping quality over all the reads at the given variant position. Instead of being a simple mean measurement, it includes the standard deviation, which is useful to asses the level of variability within the reads for that locus. Low standard deviation means that most of the data are close to the mean of the phred scaled quality score. The typical value for MQ is around 60: the filter is therefore set at rejecting any variant below 40.

- MQRankSum below -12.5; MQRankSum is the mapping quality rank sum test for mapping quality. It compares the mapping quality of the reads supporting the reference allele against the read supporting the alternate allele. A positive value shows that the reads supporting the alternate allele have higher quality than the ones supporting the reference allele. A negative value is the opposite, higher quality for the reference allele reads. If the quality scores for the reads supporting both alleles are identical the value will be 0. The default value recommended by the Broad Institute based on empirical observation is -12.5.

- ReadPosRankSum below -8.0; the read position rank sum compares the position in the reads of the reference against the position of the variant allele. This measurement is meant to check for a bias of the variant calling due to position of the alternate base in the reads. If a variant is identified by bases mainly located towards the end of the reads, there are more chances that it is due to sequencing errors as the quality of sequencing deteriorate. This deterioration of sequence quality means that bases called at the end of the read are of lower quality than bases called at the beginning of the read. If the value is negative it means that the alternate allele is found more often at the end of the read compared to the reference allele. The recommended default threshold by the Broad Institute is -8.

2.11 Variant comparison and annotation

After the variants are called and filtered the next step is to compare their frequency between the aggressive and non aggressive pools. Once a list of potential targets has been identified the variants are annotated using the most reliable source of information.

2.11.1 Comparing variants between pools

For each line and each capture the variants were compared after filtering using the following approach:

First the variant data were split into three categories:

1. Category A compared both the serial aggressors and the individuals with a family history of infanticide against the two control pools
2. Category B compared the serial aggressors against the two control pools
3. Category C compared the individuals with a family history of aggression against the two control pools

The filtered variant call files were first merged for the different pools into a unique file, and then the average frequency of each variant computed. To ensure that the means are not misrepresenting the data, a threshold of a maximum difference of 30% between the allele of pools of the same phenotypic origin was set. This threshold was chosen based on the observation of the distribution of the differences in allele frequencies between the pool, see section 2.11.4 of the results and 4.3.3 of the discussion. The first set of analysis was done using a hard threshold with the differences between the average of the pool allele frequencies for the different phenotype set to be at least 30% for category A and 50% for Category B and C. These thresholds were chosen empirically, using the number of variants passing filter as a guide to find a middle ground in the number of variant selected.

Using the hard threshold selected a number of variants but another method to define the threshold was devised, using the distribution of the difference in allele frequencies for all the variants for each of the comparisons. First the set of variants were compared to generate the difference in allele frequency, using the threshold of 30% as a cut-off for the maximum difference in allele frequency between pools belonging to the same category. This threshold was chosen based on the variability present between control pools, the majority of variants having a difference in frequency below 30% between the two control sets (see Figure 3.37 in section 3.3.2). They also needed to pass the variant call filtering in at least one of the pools to be selected as a candidate. No thresholds were set for difference between the control and infanticide pool. Instead, once generated for each category, for each line and for each capture set, the files were merged per capture set and category, giving a total of nine files, three per capture set, one per category and per capture set. The mean and standard deviation statistics of the distribution of the allele frequency difference for all variants were calculated for each file. A threshold based on the specific distribution of the allele frequency differences was chosen: mean plus three standard deviations. This should select the variants showing the highest difference in allele frequencies based on the overall allele distribution for each capture set and each category (capture 1 category A, capture 1 category B...). For more details about the choice of these thresholds, see the discussion in section 4.3.3 . Any variant that passed these criteria was classified as a candidate variant.

2.11.2 Annotation and functional analysis of the variant

This method was devised to analyse the variants using the 10.2 version of the pig genome.

Once the candidate variants were categorised, the next step was to annotate them in order to find any functional impact they could have. Due to the state of the annotation of the pig genome, comparisons with the human genome were performed in order to get a better understanding of their potential impact. The pipeline for the annotation of the variants was as follow:

1. Using the location of the pig variant, the DNA sequence around the variant was retrieved. The 300bp upstream and downstream of the variant location were taken, which produced a 600bp interval

2. The sequences were aligned against the human genome using BLAST [166] in order to identify the matching region in the human. The following parameters were used, changed from default in order to improve the search, looking for less similar sequence (between species):

   a) Match/Mismatch sequence scoring changed from 2/-3 to 1/-1. Therefore a match scored one instead of two and a mismatch is a penalty of minus one instead of minus three, making match less important and being more lenient with mismatches.

   b) Gap cost is changed: existence from 5 to 2 and extension from 2 to 1. This makes opening a gap and extending it is less costly.

   c) Searching for “somewhat similar sequences” (blastn), as pig sequences are compared to human sequences.

   d) Database used is ’Human Genomic plus Transcripts (G+T)’

3. Results were parsed to filter out weak matches, filtering criteria were:

   a) match length of 100bp between the query (pig SNP sequence) and the subject (human genomic sequence)

   b) expected value below 0.01, meaning we expect less that 1% chance of encountering a match against this database per chance.

4. Biopython was then used to annotate the parsed blast results to determine if the sequence matched in the human is in a gene, and if so, if it matches exonic or intronic sequences. The

BLAST matches were separated in 3 files:

   a) Exonic sequences

   b) Intronic sequences

   c) Not in a gene

5. For the exonic sequence, using the human coordinates and the R package biomaRt [168], the ENSEMBL biomart database of genes (human and pig) and SNPs (human) were used to annotate the data with:

   a) pig gene using the pig SNP position

   b) human syntenic gene using the pig SNP position

   c) Human SNP(s) at a position +/- 10 base pairs of the position of pig SNP in the BLAST alignment against the human.

   d) Any functional consequences linked to the human SNP(s).

6. For the intronic variants and those not in genes, biomaRt was also used but this time we looked for regulatory features using the regulatory features database to get the type of feature present (if any) and identify any genes linked to the feature. The procedure was the following:

   a) annotate the data using biomaRt to get the regulatory feature ID linked to the human region in synteny with the pig genome

   b) get all the genes linked to that regulatory feature using GeneCard (www.genecards.org [177, 178])

   c) Using DAVID (https://david.ncifcrf.gov/ [179]), the genes linked to the feature are curated for any interesting traits or functions.

2.11.3 Updating coordinates to *Sus scrofa* genome release 11

During the later part of this work a new assembly of the *Sus scrofa* genome (SS 11) was released (see discussion 4.2.5). The genomic positions of variants and capture set probes needed to be updated to the new build coordinates. It is important to check each candidate variant genomic position against the new build in order to look at potential new consequences for the region around it. The capture probe coordinates were checked to make sure the target region was still the same as its original design. A pipeline was designed to process the capture set data and the candidate variant using the biomaRt package [169, 168], the Ensembl compara PERL API [152] and various python scripts to handle the data. The Ensembl compara API was used to retrieve the human genomic coordinates from the pig genomic coordinate, allowing pig genomic coordinates to be mapped to human ones.

This pipeline was applied to the first set of variant selected, using hard thresholds. Given the relatively small number of variants selected using this data, it was necessary to try to annotate most of them. For the variants selected using the threshold based on the distribution of allele frequency differences, a simpler version of this pipeline was used (see next section 2.11.4).

The pipeline to update the candidate variant positions is as follows, in total 9 tables were put through the pipeline, 3 for each capture set corresponding to each category (A, B or C):

1. The variants were processed using biomaRt to update their coordinates to SS 11 and find any human homologues:
    a) The RefSNP or RS id was queried against biomart to get the updated position in SS 11 if it is available
    b) If no position was returned to the genomic coordinates of the variant are used to query for an updated RS id and the coordinate in SS 11
    c) If there was still no match the variant is flagged for follow up with another method (see below)
    d) If new coordinates were found, they were used to query for any human homologous gene at the SS 11 position.
2. Variants that did not update to SS 11 were submitted to the NCBI remap tool [180] in order to update the coordinates
    a) When several possible locations were obtained, the most likely location was curated manually by looking at the chromosome it mapped to, how similar in size was the mapped interval to the size of the region targetted
    b) Once a multiple locations entry had been resolved, the flagged candidate variant followed the same step as the other variants.
3. Using the pig variant coordinates the Ensembl compara perl API was used to get the corresponding human syntenic blocks. Any variant that did not return any human coordinates still went through the next steps but was run separately from variants with human coordinates.
4. For candidate variants updated with the compara API, the synteny coordinates were used to query biomart using biomaRt to get any human gene mapping to that location
5. The SSC11 coordinates were used to get any pig genes mapping to that location.
6. The R package dplyr was used to merge the variants that had to be updated separately with the other variants, creating an updated version of the input file.

The pipeline to update the capture coordinates was very similar to what had been done to the candidate variants. However there were some notable differences because the data were genomic intervals rather than a single genomic location.

1. The files provided by Agilent gave coordinates for the probes covering a targetted region. Using R dplyr the coordinates were grouped according to the region targetted, taking the minimum and maximum coordinates of the probes for that region.

2. Contrary to the variant data, the capture region could not be updated efficiently using biomart, so the NCBI remap tool was used to update the coordinate to SSC11:

   a) When several possible locations were obtained, the most likely location was curated manually. This was done by looking at the chromosome it mapped to, and how close was the mapped interval in relation to the size of the region targetted.

   b) Once multiple locations entry had been resolved, the capture regions followed the same steps as the rest of the capture regions.

3. Using the updated region coordinates, the Ensembl compara perl API was used to get the corresponding human syntenic blocks. Any region that did not return any human coordinates will still go through the next steps

4. Given the size of the target regions, the API returned multiple blocks for each region submitted:

   a) The data were first sorted by genomic location and a cluster id was added to each line. If the chromosomes were the same as the previous line and the difference between the end of the previous block and the current one is less than 10000 bases, they were attributed the same cluster id, otherwise a new id was generated.

   b) The lines with the same cluster id were merged using dplyr and taking the minimum and maximum coordinates for each entry with the same cluster id.

5. For candidate regions updated with the compara API, the synteny coordinates were used to query biomart using biomaRt to get any human genes mapping to that location

6. For regions that have been updated to SSC11, the updated coordinates were use to get any pig genes mapping to that location.

7. To check if the coordinates had changed dramatically the Gviz package was used to generate a genomic region plot of:

   a) The targeted region on SSC10.2 (original capture coordinates)

   b) The targeted region on SSC11 (updated coordinates)

   c) The targeted region corresponding to the syntenic region on the human genome.

8. Comment is added to the file once the region has been checked against the genomic region plot

Finally the variant data was checked against the updated capture set to check if any candidate variants were no longer within the targetted region.

2.11.4 Curation of SNPs and regions of interest

After the design of a new filtering method (see section 2.11.1), a shorter version of the pipeline above was used to simply update the SNPs to SSC11 using biomaRt and at the same time get the SIFT prediction for the impact of the SNPs [181]. SIFT is an algorithm designed to identify potential deleterious change in the amino acid (AA) chain of the protein, caused by a SNP allele. The individual SNPs selected were SNP classed as missense SNP, causing a potential change in AA at this location. SIFT can predict if the effect might be deleterious but the prediction has a level of uncertainty. As only a few SNPs were selected as missense, a more in depth analysis was required. In order to check if there is a potential effect on the protein structure and function, the following pipeline was applied to candidate SNP.

1. SNP was queried using the Ensembl website

2. The alignment against other species (phylogenetic context) was used to check the consensus sequences for the same gene for other species. The phylogenetic context displayed the alignment of the SNP and the region around it against other species.

3. If the gene was unknown, the region comparison tool was used with the Human and Mouse genomes in order to see if a known gene is present at the corresponding location in these genomes

4. To better estimate the impact of the of the amino acid change, the protein sequence around the target amino acid was blasted against the protein database using the NCBI website, using the default setting for blasting protein against protein.

5. The blast results were used to assess if the region of the protein is highly conserved or not.

6. The impact of the amino acid change was also evaluated by looking at their properties, notably their charge, polarity and hydrophobic or hydrophilic nature.

For the regions where a large number of SNPs were identified, the following pipeline was used:

1. The regions and SNPs were plotted using the version 11 of the genome in order to identify genes in their vicinity.

2. If no genes are presents, the region was queried in Ensembl for region comparisons with the Human and the mouse.

3. Any gene covered by the SNPs were queried in GeneCards [177] to look for any interesting disorder or function,

4. The gene was also queried in the NCBI gene database, for both human and mouse to learn more about their function

3 Results

3.1 Genotyping microarray results

3.1.1 Summary of the data available

Two sets of data were available for this study, one was generated in Cambridge and the other one provided by Genus PLC. Tables 3.1 and 3.2 show the data available for each set. The sets were merged to form one unique dataset that was then filtered in order to perform the FBAT and PO tests.

Lines	Control samples (females)	Infanticide samples	No status (female)	Male samples	Female samples	Total
Line B	178	122	1	8	301	309
Line C	110	101	1	7	212	219
Line D	210	159	3	51	372	423
Line H	282	189	1	6	472	478

Table 3.1: Samples typed in Cambridge for the four lines, not status females are female with no entry for the infanticide status.

Lines	Control samples (females)	Infanticide samples	No status (female)	Male samples	Female samples	Total
Line B	104	17	3	84	124	208
Line C	72	20	1	96	93	189
Line D	-	-	-	-	-	-
Line H	0	0	0	150	0	150

Table 3.2: Sample typed by Genus for the four lines, no status females are female with no entry for the infanticide status.

3.1.2 Quality control

Using the pedigree information the data were filtered to leave only the families with at least two individuals, one of which must be an infanticide sow. The QC was then run on the data using the parameters described in section 2.5 using PLINK. The summary of this QC is given in table 3.3. Another QC step performed was to look at the Minor Allele Frequency (MAF). Table 3.4 summarises the low MAF SNPs in the dataset, displaying the percentage of 0% MAF (homozygous SNP) and below 5% MAF (low frequency variant), up to a quarter of the SNP typed have low MAF (below 5%), making them low frequency variants for these lines. After QC the data were loaded in FBAT and PLINK for the PO analysis. FBAT returned some errors after the heritability checks within each family. Some families had to be removed from the analysis as the number of errors proved too important, suggesting that the individual in these family might not be related. This can be caused by cross fostering, when a piglet has moved to another dam before it could be recorded properly. Between 2 and 4 families were removed from each line, the details of the families analysed in FBAT are given in table 3.5.

Lines	Markers	Samples	Call rate	HWE filter	Genotype missing (95%)	Missing individual filter (90%)	Passed QC markers
Line B	61062	206	96.80%	22	4473	0	56567
Line C	61062	158	95.71%	31	4251	0	56780
Line D	61062	188	95.83%	28	3915	0	57119
Line H	61062	297	95.80%	36	4342	3	56684

Table 3.3: Markers information for Family Based Association and Parent of Origin: HWE filters stands for Hardy Weinberg filter which tests for the Hardy Weinberg equilibrium, The genotype missing is the threshold for removing markers for which not all the individuals have a call (5% of individual missing), the missing individual filter is to remove individuals for which the overall call rate is below the threshold (10% or more missing).

Line	Total Markers	0% MAF	Below 5% MAF	Percentage 0%	Percentage 5%
B	56567	6325	14743	11.18	26.06
C	56780	9375	14628	16.51	25.76
D	57119	6779	12329	11.87	21.58
H	56684	6825	13085	12.04	23.08

Table 3.4: Table summarising low Minor Allele Frequency (MAF, the lowest allele frequency from each SNP), number of SNPs with 0% MAF (homozygous SNPs) and less the 5% MAF.

Lines	Total samples	Families	Families with at least 3 members	Families removed	Infanticide females	Control females	Males	Females with no phenotypic status
Line B	206	76	45	4	83	76	47	0
Line C	158	50	32	2	73	44	40	1
Line D	188	56	33	3	85	62	41	0
Line H	294	97	63	3	122	104	68	0

Table 3.5: Samples for family based association. Total number of samples represents the number of samples belonging to a family. Families represent the total number of families in the set. Families with at least 3 members represent families when more individuals have been typed than the minimum of 2 (one parents can be missing). Family removed: the number of family excluded due to heritability errors. Infanticide/Control female, the number of Infanticide and Control females present in the families tested. Male: number of sires present in the set. Females with no phenotypic status: females for which the aggressive status was unknown.

The QC of the genotyping data is good (see table 3.3) as the average sample call rate is above 95% and only a few samples failed to meet our filtering threshold of 95% call rate. Between 3% and 8% of the samples were filtered out at this stage. The Hardy-Weinberg test removed a small proportion of the markers (less than one percent for all the samples). However the “missing genotypes” filtering removed around 7% of the markers for each line as they were typed in less than 5% of the samples. This figure is consistent across the different lines studied, therefore it is unlikely to be line specific, it is more likely to reflect the quality of the markers present on the array. There is a strong possibility that the markers removed did not get typed successfully because the design was based on a less refined version of the genome. Therefore the target sequences might not be accurate, resulting in systemic failure when typing these markers. After filtering we were left with more than 92% of the markers for the analysis.

For the FBAT and PO analysis the individual maximum percentage of missing SNP was set to 10%: i.e, any individual with more than 10% of the SNP calls missing was excluded. This threshold was chosen in order to not be too stringent, and retain some individuals with some SNP missing. The aim was to have the highest possible number of individuals per families for analysis. Furthermore because of the nature of the genome used to design the array (*Sus scrofa* 7) and the breeds selected for its design; there is a chance that some of our lines will not match the reference breed. In our analyses we have a mixture of pig breeds and lines (Duroc, Large White and Landrace), which could cause some mismatch with the array. Only line H has lost individuals due to this filter (table 3.3). The failure of these 3 samples is most likely due to the DNA sample quality.

Using biomart it is possible to check how many markers are still valid with the latest build of the genome (table 3.6). For each line more than 87% of the SNPs are still found using *Sus scrofa* genome build 11 (SS 11). Some of the SNP missing from the cross-referencing might because the SNP identifier has changed and is not being recognised as valid. Some SNPs might also have been invalidated. In any case, around 50,000 SNPs are left to test in this data set for each of our lines.

	Input SNP	Valid in Biomart	Percentage total
All array	61565	53855	87.5%
Line B	56567	49609	87.7%
Line C	56780	49789	87.7%
Line D	57119	50086	87.7%
Line H	56780	49789	87.7%

Table 3.6: SNP left after QC, used as a Biomart query and still present using version 11 of the pig genome.

3.1.3 Pig array 60k from Illumina

For this study the genotyping data comes from 2 sources: the porcine 60k version 1 array from Illumina (Genus PLC data and first run of data from Cambridge) and the version 2 of the porcine 60k from Illumina (second run of Cambridge data). The data from both sets were merged using the common SNPs between the two arrays in order to generate the data set used for both set of analysis. The SNPs remaining were therefore mostly those designed for version 1 of the array. The 60k version 1 array was constructed in 2009 and the state of the pig genome at this point (version 7) was very fragmented. While the design of the array described in [98] used the best technology available at the time, there are several potential problems with the approaches taken, due to the limitations of the technology available at the time. The reads used for the design were generated from five different pools of animals, using very short sequencing length (36bp) and low quality thresholds to identify polymorphic sites . Reads with a quality score above 10 were considered good enough to be used, whereas nowadays a minimum quality score of 20 is recommended. Furthermore, the reference genome used for the alignments was only sequenced at an average depth of 4 reads per bases, there is no mention of the depth of coverage used for the selection of SNPs. Despite these drawbacks, the candidate SNPs were curated and passed through several rounds of selection, which should improve their quality. Some issues are still left after selection and visible when looking at the minor allele frequency (MAF) of the SNPs on the array for the different lines. As shown on table 3.4, there are between 11 and 16% of the SNPs that are homozygous on the array. Furthermore, around a quarter of the SNPs have a MAF below 5%. This means the minor allele frequency is very low for these SNPs in the lines studied, raising questions about their usefulness to represent LD blocks. This will likely lower the numbers of LD blocks analysed with this array for this thesis. It is expected that large gaps between SNPs will be present and some regions of the genome will not be well covered. This can result in some of the LD blocks not being typed properly, or missed entirely, although one mitigating factor is that in some of the breeds studied the LD blocks appeared to be conserved over large intervals [182].

While the array has many potential drawbacks, it was still at the time of this study the best method to perform a large scale investigation of the genetic causes of maternal infanticide. There was no alternative other than typing each SNP individually, a much more expensive option.

3.1.4 FBAT and PO test and significance thresholds.

Both tests were chosen due to the addition of data provided by Genus PLC. This additional data completed families and pedigrees already available in the previous dataset from Quilter et al [2]. The FBAT test is more appropriate for our dataset than a traditional association test as our data is composed partially of families. The approach using unrelated individuals for a association test was already done in Quilter et al [2]. Using another approach to query the data is a good way of confirming some of the results already found. Given that we had a good number of families, using tests designed for this type of data was a logical step. Using the relationships present, it was expected that we would have more power for the analysis. Another approach that we wanted to test on this type of data is the preferential transmission of allele from mother to daughter. It was shown that this might be a factor in the heritability of the infanticide trait [3]. In order to investigate this we used the PO test and concentrated on the dam to daughter preferential transmission of alleles.

The number of families and samples selected for the FBAT and PO analysis are given in table 3.5 of section 3.1.2. The number of families available for each line is between 50 and 97: two lines (line C and D) have around 50 families while line B has 76 and line H has 96. When performing the test, FBAT returns the number of families for which the association in the presence of linkage is found. For the PO test, the ratio of transmitted allele (transmitted over un-transmitted allele) is also given. This can be used to assess the likelihood of the results being of interest, if a large number of families supports the evidence for association, or, if the ratio of the transmitted over the un-transmitted allele is either large or very small.

The aim of both of these tests is to identify or confirm regions of the genome linked to maternal infanticide, in order to use sequence capture to investigate these regions in more details and narrow down variants. The variants identified will be used to refine genes or regions of interest that might play a key role in the maternal infanticide phenotype. There are several reasons why it would be detrimental to be too stringent when filtering the results of the FBAT and PO tests. To begin with, the phenotype of interest exerts little selection pressure on the animals affected, these animals come from lines generated for their high breeding values and quality traits (e.g. meat and reproductive trait). Such high value animals are unlikely to be culled unless they are extreme cases. Furthermore it is also known from previous studies [2, 17, 80, 14], that several regions of the genome are affected and multiple variants in the genome might be linked to this phenotype. It is unlikely that the penetrance of any of these variants or regions linked to this phenotype will be high, our trait of interest appears to fit the classic, multiple components phenotype. Therefore being too stringent while filtering the results could remove regions of interest. The FBAT and PO tests are therefore used primarily as a screening tool.

The threshold chosen for the FBAT analysis is $-log_{10}(pvalue) > 4$ which corresponds to a p-value of 1.10^{-4}. There is no permutation or adjustment p-value available for the FBAT test unless the pedigree approach is used. This threshold >4 is below the genome wide significance level usually used in GWAS and genome wide studies, which is 5.10^{-8}or a $-log_{10}(p - value) > 7.3$. For the PO test the threshold was set at 0.03, which will result in a ratio of untransmitted over transmitted or transmitted over untransmitted of at least 3/1. This will again identifies a variety of regions that were targetted for the sequence capture.

3.1.5 Family Based Association

The Family Based Association Test (FBAT) was used in order to test for association in the presence of linkage using families present in the data set. Using the family structure present in the dataset made it a different and complementary approach from the previous study which looked at non related samples.

The results of the family based association study are as follow:

- For line B, 15 SNPs reached a significance of $10^{-5}(-log_{10}(p - value) > 5)$, on 9 different chromosomes: chromosome 1, 2, 3, 5, 7, 10, 14, 15 and 17, see table 3.7.

- For line C, 8 SNPs reached a significance of $10^{-5}(-log_{10}(p - value) > 5)$, on 5 different chromosomes: chromosome 2, 7, 10, 15 and 18, see table 3.8.

- For line D, 29 SNPs reached a significance of $10^{-5}(-log_{10}(p - value) > 5)$, on 12 different chromosomes: chromosomes 1, 2, 3, 5, 6, 7, 8, 9, 10, 14, 15, 17, see table 3.9.

- For line H, 41 SNPs reached a significance of $10^{-5}(-log_{10}(p-value) > 5)$, on 14 different chromosomes, chromosome 1, 2, 3, 4, 6, 7, 8, 9, 10, 13, 14, 15, 16, 18, see table 3.10.

Figures 3.1, 3.3, 3.5 and 3.7 shows the Manhattan plots for each of the lines (B, C, D and H). Figures 3.2, 3.4, 3.6 and 3.8 are the QQ plot for the FBAT p-value for each line. Table 3.11 displays the lambda value for the QQ plot for the FBAT results. The lamda values are relatively close to one expected for line H, for which some stratification might be present as the value is above 1.1.

Marker	RS Identity	Chromosome	Position SS 11	Number of s informative families	p-value	$-log_{10}(p-value)$
ASGA0005756	rs80837832	1	219,114,187	43	5.60E-11	10.25
ALGA0102837	rs81329722	2	39,427,398	43	5.60E-11	10.25
MARC0009481	rs81257081	3	72,010,079	26	2.60E-06	5.59
M1GA0007286	rs81383309	5	5,672,567	41	1.55E-10	9.8
MARC0055096	rs80880714	7	9,057,028	36	2.02E-09	8.69
ALGA0039425	rs80873322	7	21,554,498	20	4.22E-06	5.37
BGIS0000098	-	7	NA	37	5.35E-09	8.27
ALGA0118632	rs81325586	10	14,208,472	25	5.73E-07	6.24
H3GA0033419	rs81439383	12	9,767,814	31	9.55E-06	5.01
ASGA0066525	rs80813041	14	123,010,045	36	2.02E-09	8.69
ALGA0085872	rs81453527	15	75,057,741	42	1.49E-10	9.83
ALGA0086085	rs80921770	15	79,583,768	37	1.21E-09	8.92
H3GA0044934	rs80926252	15	119,410,395	36	2.02E-09	8.69
MARC0083113	rs81266760	18	37,977,138	42	9.32E-11	10.03
ASGA0089892	rs81308090	18	39,430,154	43	5.60E-11	10.25

Table 3.7: FBAT line B results table. The Marker column show the Pig consortium ID used by Illumina in their annotation file. The RS identity is the dbsnp id for the marker. Chromosome and position SS11 give the pig chromosome and the position after remapping the position in *Sus Scrofa* 10.2 to *Sus Scrofa* 11. The number of informative families shows the number of families that contributed to the test for that marker. Finally the last two columns show the p-value and $-log_{10}(p-value)$ for the FBAT test.

Marker	RS Identity	Chromosome	Position SS 11	Number of informative families	p-value	$-log_{10}(p-value)$
ALGA0102837	rs81329722	2	39,427,398	36	5.59E-11	10.25
H3GA0022731	rs80910844	7	100,110,224	30	2.43E-08	7.61
ALGA0118632	rs81325586	10	14,208,472	30	1.21E-09	8.91
ALGA0100690	rs81477669	10	45,783,989	27	6.03E-06	5.21
H3GA0044934	rs80926252	15	119,410,395	36	5.59E-11	10.25
MARC0083113	rs81266760	18	37,977,138	36	5.59E-11	10.25
ASGA0089892	rs81308090	18	39,430,154	36	5.59E-11	10.25
BGIS0000098	-	7		30	$9.76E-6$	5

Table 3.8: FBAT line C results table. The Marker column show the Pig consortium ID used by Illumina in their annotation file. The RS identity is the dbsnp id for the marker. Chromosome and position SS11 give the pig chromosome and the position after remapping the position in *Sus Scrofa* 10.2 to *Sus Scrofa* 11. The number of informative families shows the number of families that contributed to the test for that marker. Finally the last two columns show the p-value and $-log_{10}(p-value)$ for the FBAT test.

Marker	RS Identity	Chromosome	Position SS 11	Number of informative families	p-value	$-log_{10}(p-value)$
ASGA0001011	rs80870251	1	10,968,552	27	5.73E-07	6.24
ASGA0004231	rs80794948	1	110,239,011	15	2.73E-06	5.56
ALGA0005867	rs80914010	1	125,222,441	34	2.05E-12	11.69
ALGA0114601	rs81343604	1	NA	16	7.10E-06	5.15
ASGA0005756	rs80837832	1	219,114,187	40	3.35E-11	10.47
DRGA0002016	rs80904152	1	223,861,733	22	1.50E-06	5.82
INRA0007489	rs319558321	1	262,121,588	19	1.50E-06	5.82
H3GA0006383	rs81356441	2	25,215,306	31	1.77E-09	8.75
ALGA0102837	rs81329722	2	39,427,398	40	3.35E-11	10.47
ALGA0018634	rs80784123	3	44,000,093	29	2.93E-09	8.53
M1GA0007286	rs81383309	5	5,672,567	40	3.35E-11	10.47
MARC0009578	rs81258152	6	120,053,311	16	4.22E-06	5.37
H3GA0018823	rs81391651	6	133,185,154	19	5.74E-06	5.24
H3GA0022731	rs80910844	7	100,110,224	34	2.96E-07	6.53
MARC0032237	rs80785162	8	111,171,988	26	2.28E-09	8.64
ALGA0054131	rs81414030	9	86,593,109	13	7.74E-06	5.11
ALGA0118632	rs81325586	10	14,208,472	40	3.35E-11	10.47
ALGA0100690	rs81477669	10	45,783,989	36	2.26E-12	11.65
MARC0016782	rs80799328	14	14,414,380	24	9.63E-07	6.02
ASGA0066525	rs80813041	14	123,010,045	40	3.35E-11	10.47
MARC0101508	rs81278062	15	10,585,538	25	1.14E-07	6.94
ALGA0085872	rs81453527	15	75,057,741	35	2.38E-06	5.62
ALGA0086085	rs80921770	15	79,583,768	40	3.35E-11	10.47
H3GA0044934	rs80926252	15	119,410,395	40	3.35E-11	10.47
MARC0039970	rs81233198	18	25,273,873	15	1.60E-06	5.79
MARC0083113	rs81266760	18	37,977,138	40	3.35E-11	10.47
ASGA0089892	rs81308090	18	39,430,154	40	3.35E-11	10.47
M1GA0023271	rs81470243	18	45,399,121	18	2.52E-06	5.6
H3GA0047065	rs81461972	-	NA	35	7.55E-13	12.12

Table 3.9: FBAT line D results table. The Marker column show the Pig consortium ID used by Illumina in their annotation file. The RS identity is the dbsnp id for the marker. Chromosome and position SS11 give the pig chromosome and the position after remapping the position in *Sus Scrofa* 10.2 to *Sus Scrofa* 11. The number of informative families shows the number of families that contributed to the test for that marker. Finally the last two columns show the p-value and $-log_{10}(p-value)$ for the FBAT test.

Marker	RS Identity	Chromosome	Position SS 11	Number of informative families	p-value	$-log_{10}(p-value)$
ASGA0001011	rs80870251	1	10968552	24	4.46E-07	6.35
ALGA0005867	rs80914010	1	125222441	52	7.52E-06	5.12
ALGA0114601	rs81343604	1	NA	33	7.24E-10	9.14
DRGA0002016	rs80904152	1	223861733	20	2.73E-06	5.56
INRA0007489	rs319558321	1	262121588	24	7.24E-08	7.14
ALGA0102837	rs81329722	2	39427398	71	4.76E-18	17.32
ALGA0018634	rs80784123	3	44000093	39	5.62E-08	7.25
DIAS0000481	rs343661863	4	15671825	28	1.24E-06	5.91
ALGA0027584	rs81380005	4	103968409	22	3.41E-07	6.47
H3GA0052806	rs81334603	6	170305876	19	7.74E-06	5.11
H3GA0019527	rs80949107	7	3676156	37	4.33E-10	9.36
ALGA0038342	rs80872016	7	7273490	21	7.10E-06	5.15
ALGA0039425	rs80873322	7	21554498	23	5.73E-07	6.24
MARC0049081	rs80824189	7	73999704	27	9.46E-09	8.02
BGIS0000098	-	7		58	8.70E-06	5.06
M1GA0010553	rs80833324	7	88763190	39	1.52E-06	5.82
H3GA0022731	rs80910844	7	100110224	57	2.34E-10	9.63
DIAS0001763	-	8	NA	32	5.57E-06	5.25
ALGA0054131	rs81414030	9	86593109	18	1.62E-06	5.79
ALGA0118632	rs81325586	10	14208472	45	4.34E-12	11.36
H3GA0029984	rs81423899	10	38596942	38	3.02E-10	9.52
ALGA0100690	rs81477669	10	45783989	56	1.42E-21	20.85
DRGA0011972	rs80797698	13	9952427	60	3.78E-07	6.42
ALGA0069762	rs80911538	13	41069431	21	2.03E-07	6.69
ALGA0073309	rs81441892	13	185104190	24	3.38E-09	8.47
MARC0066830	rs81255192	13	192073106	20	4.59E-06	5.34
ALGA0074022	rs81443003	13	204827712	25	2.03E-07	6.69
MARC0031872	rs80919390	14	2704204	22	5.74E-06	5.24
ALGA0078467	rs80855882	14	65655137	56	2.98E-06	5.53
H3GA0042218	rs80915578	14	119983231	24	2.11E-08	7.68
MARC0101508	rs81278062	15	10585538	27	3.19E-07	6.5
ASGA0068898	rs81451733	15	19720520	30	3.38E-09	8.47
MARC0046889	rs81238797	15	39419144	20	2.73E-06	5.56
H3GA0044934	rs80926252	15	119410395	69	1.31E-17	16.88
ALGA0090556	rs81459227	16	44602169	31	1.41E-06	5.85
MARC0039970	rs81233198	18	25273873	34	1.21E-09	8.92
MARC0083113	rs81266760	18	37977138	71	4.76E-18	17.32
ASGA0089892	rs81308090	18	39430154	69	1.31E-17	16.88
MARC0002018	rs80830421	-	233795145	35	1.45E-07	6.84
MARC0026936	rs81293197	-	NA	31	1.50E-08	7.82
H3GA0047065	rs81461972	-	NA	20	2.10E-06	5.68

Table 3.10: FBAT line H results table. The Marker column show the Pig consortium ID used by Illumina in their annotation file. The RS identity is the dbsnp id for the marker. Chromosome and position SS11 give the pig chromosome and the position after remapping the position in *Sus Scrofa* 10.2 to *Sus Scrofa* 11. The number of informative families shows the number of families that contributed to the test for that marker. Finally the last two columns show the p-value and $-log_{10}(p-value)$ for the FBAT test.

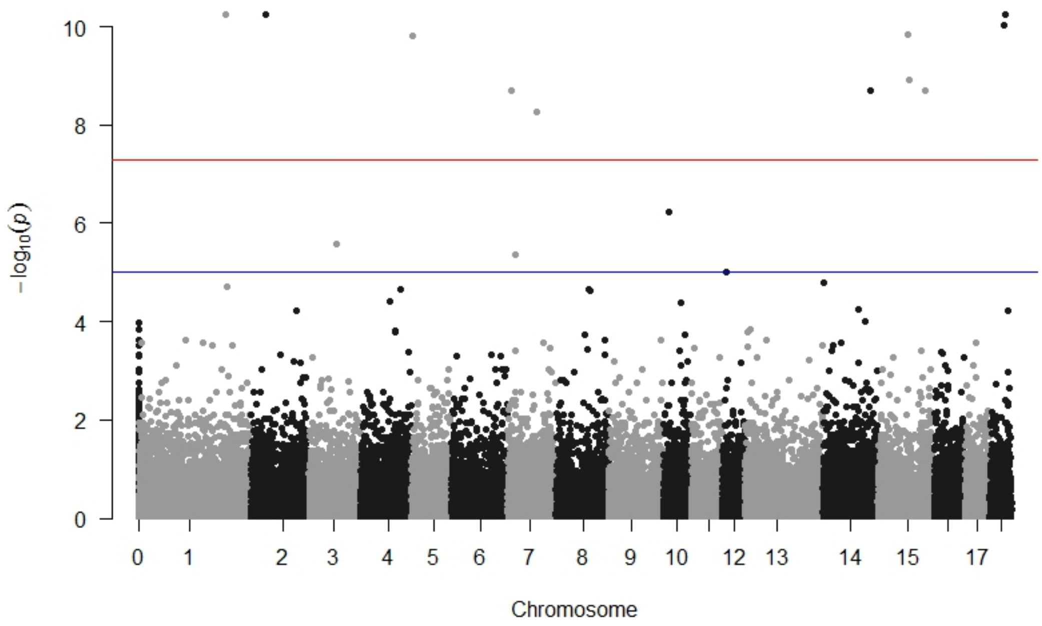


Figure 3.1: Manhattan plot of the FBAT results for line B. The blue line shows the chosen threshold, the red line is genome wide significance.

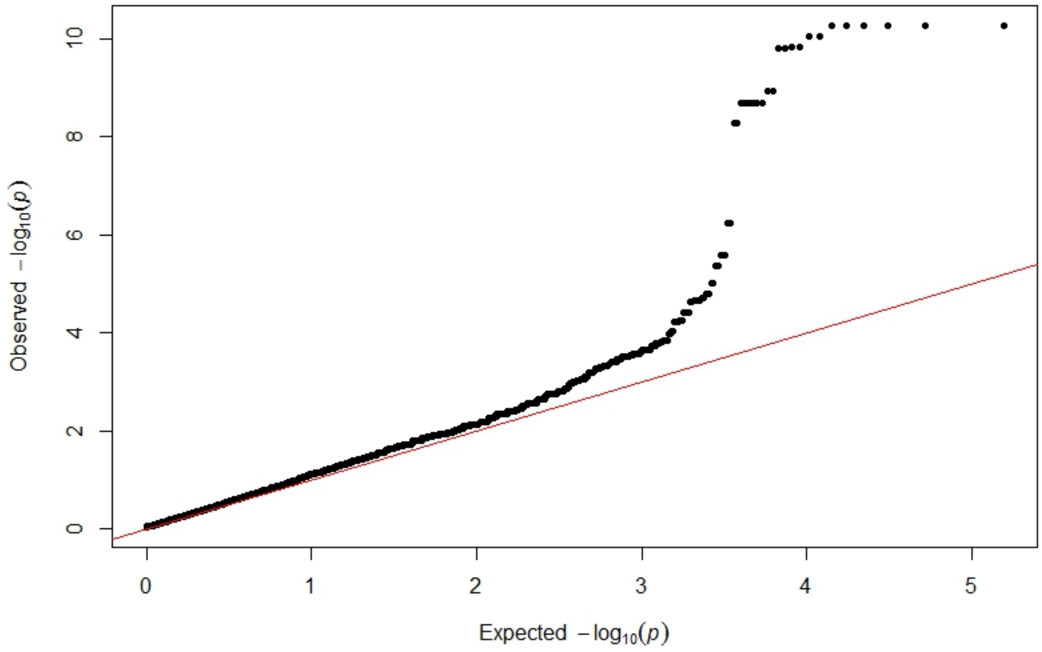


Figure 3.2: QQ plot for the p-value of the FBAT test, for line B. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 1.034.

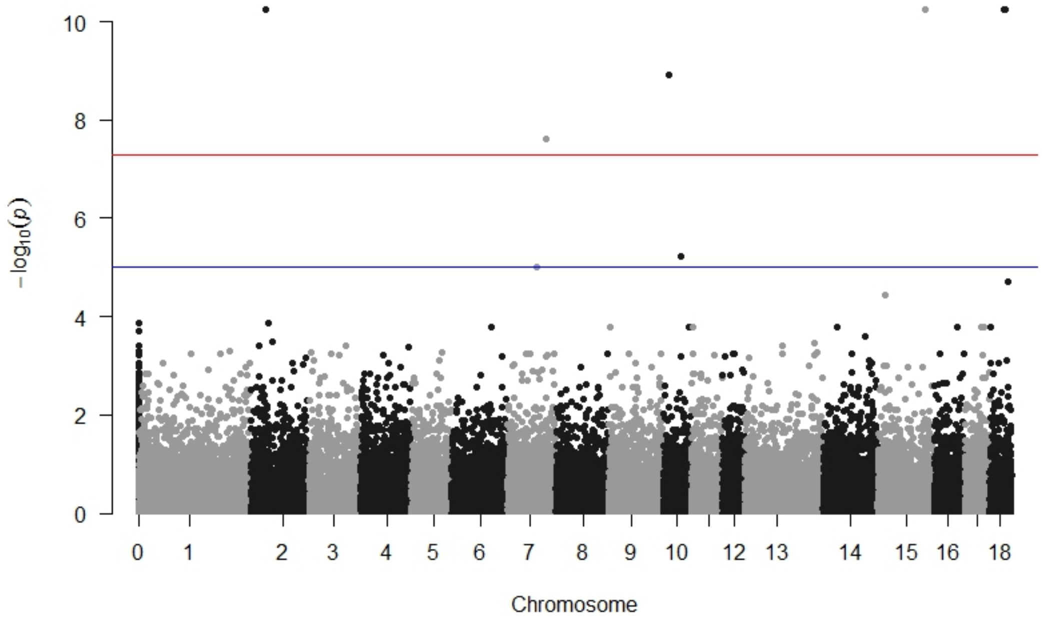


Figure 3.3: Manhattan plot of the FBAT results for line C. The blue line shows the chosen threshold, the red line is genome wide significance.

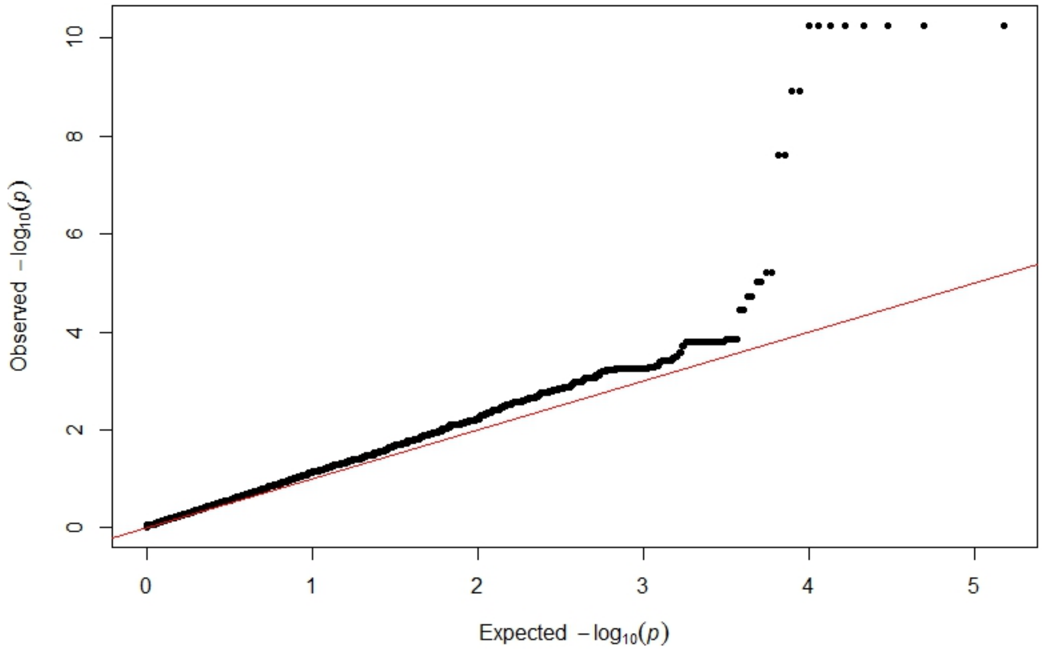


Figure 3.4: QQ for the p-value of the FBAT test, plot for line C. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 1.099.

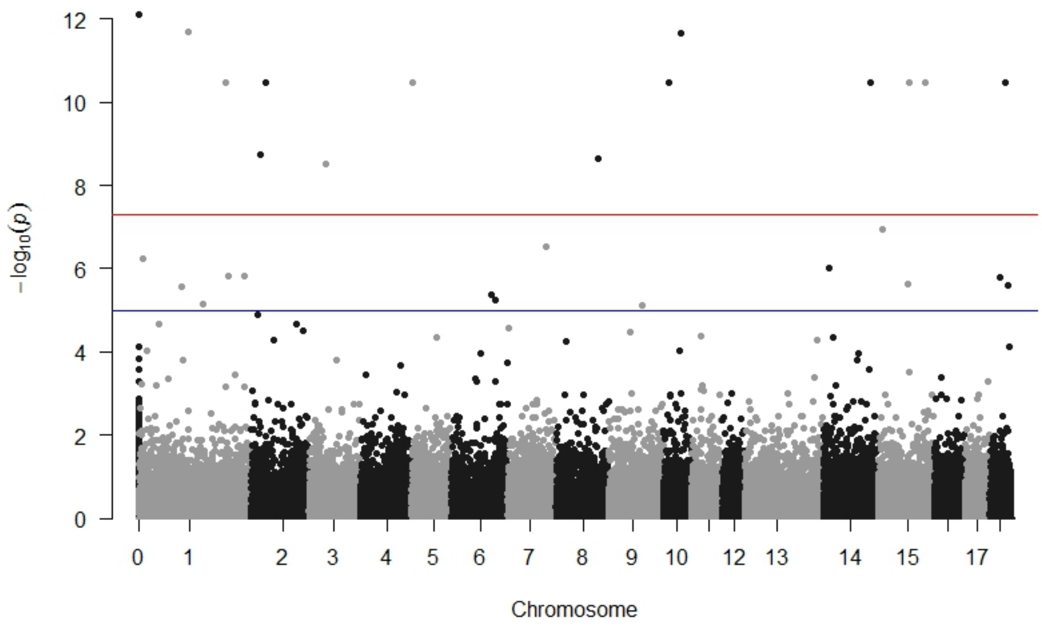


Figure 3.5: Manhattan plot of the FBAT results for line D. The blue line shows the chosen threshold, the red line is genome wide significance.

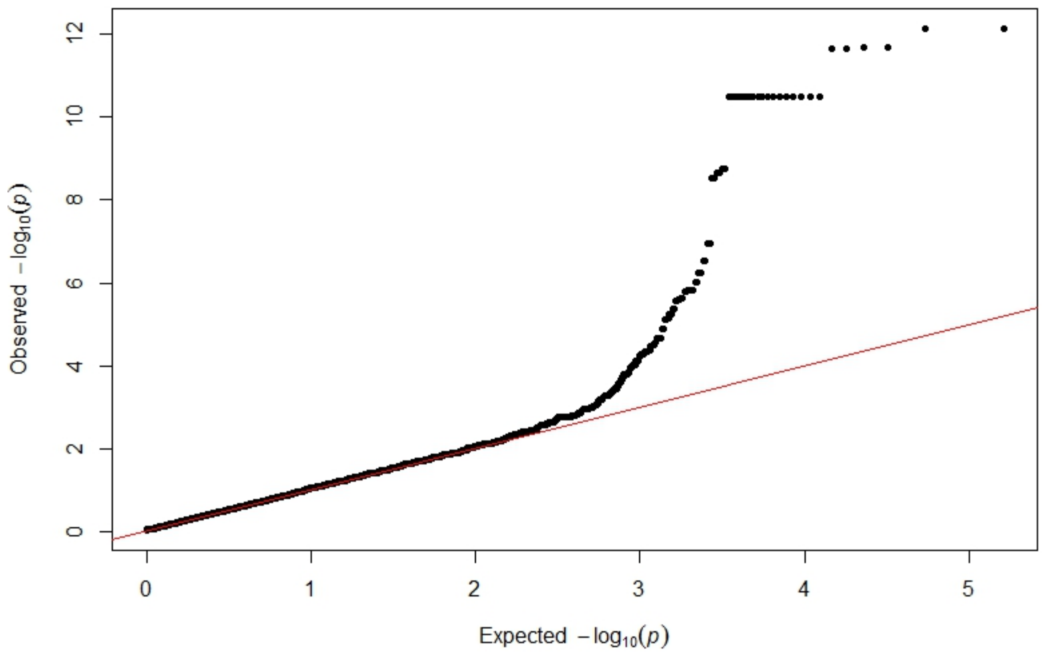


Figure 3.6: QQ plot for the p-value of the FBAT test, line D. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 1.034.

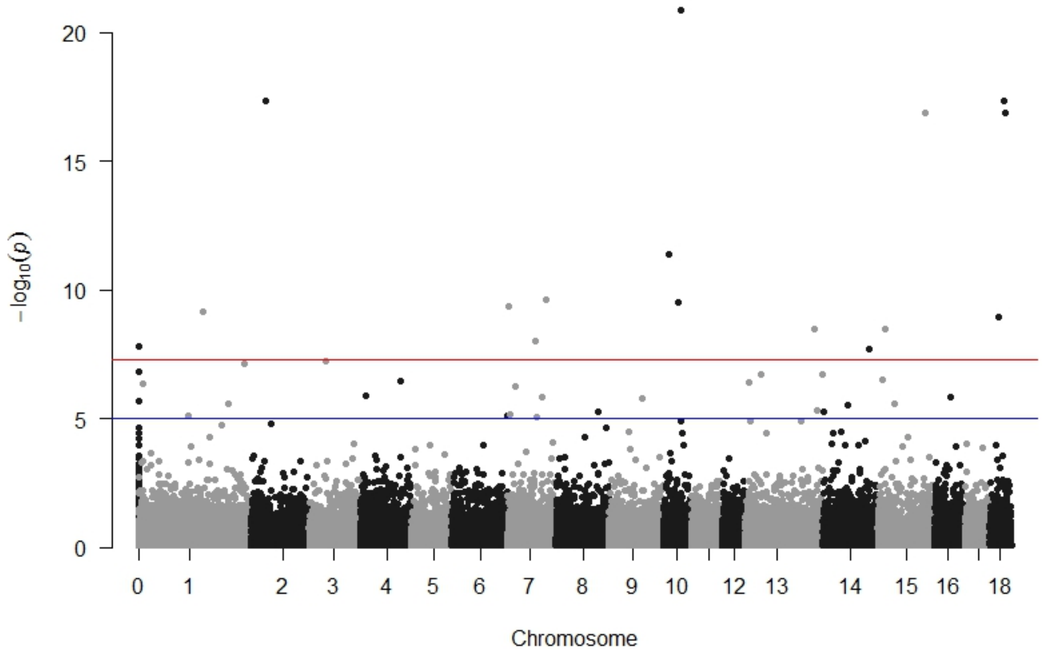


Figure 3.7: Manhattan plot of the FBAT results for line H. The blue line shows the chosen threshold, the red line is genome wide significance.

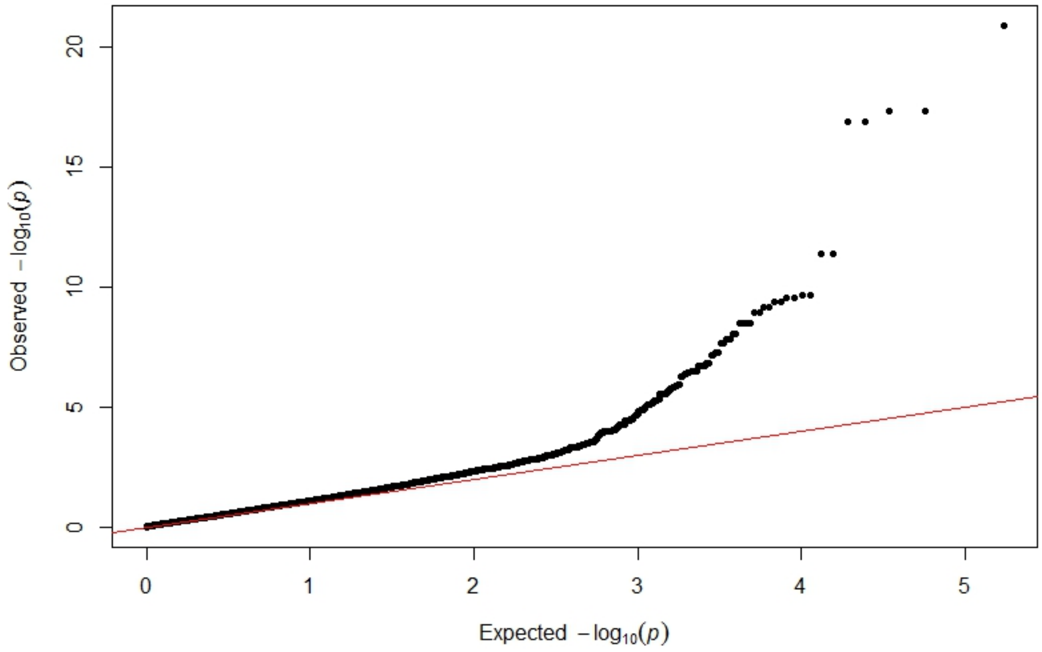


Figure 3.8: QQ plot for the p-value of the FBAT test, for line H. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 1.121.

Line	Lambda value
B	1.034
C	1.099
D	1.034
H	1.121

Table 3.11: Lambda value for the p-value distribution of the FBAT results for all the lines.

Using the threshold discussed in section 3.1.4, a small number of SNPs were identified as significant on a number of chromosomes. For lines B and C only a few SNPs, 15 and 8, were significant, while a larger number of SNPs were significant for lines D and H, 29 and 41 respectively. These results fit well with the incidence of maternal infanticide in the lines studied, lines B and C having the lowest incidence of maternal infanticide with around 5%, (4.8% and 5.9% for line B and C respectively) while lines D and H have a higher incidence of maternal infanticide with 10.8% and 10.3% respectively. These results do not correlate with the number of families present in each lines (see table 3.5 in section 3.1.2). Lines B and H have the largest number of families present with 76 and 97 families respectively. Lines C and D have a lower number of families with 50 and 56 respectively. There might be an impact from the number of families with at least 3 members, as line H has the highest number of families with more than three members: 63. However FBAT can infer missing genotypes if one of the parents is absent, using heterozygote SNPs in the offspring, which can mitigate some of the missing data for some of the markers. The numbers of informative families for the significant SNPs are fairly similar for lines B and C (table 3.7 and 3.8), varying between 20 and 40. For line D, the numbers of informative families is between 13 and 40 per significant intervals (see table 3.9). Finally for line H the number of informative families is between 71 and 19 (see table 3.10). There is no obvious corelation between these numbers and the number of significant results.

For all lines there are SNPs that reach genome wide significance: 11 for line B, 6 line C, 15 for line D and 17 for line H. These make up almost half of the total number of significant SNPs passing our threshold. Again the numbers are correlated with the incidence of maternal infanticide in the lines.

The “Manhattan plots” are used in genome wide genetics studies to show the regions of each chromosome where significant SNPs are grouped, creating a peak of significant SNPs. The “Manhattan plots” for the FBAT results are displayed in section 2.7, figure 3.1, 3.3, 3.5 and 3.7. There was no group of significant SNPs per chromosome, as might be expected from a GWAS analysis. Only single SNPs were reaching our threshold or genome wide significance, which might be caused by several factors. First, our phenotype of interest is probably linked to several genes across the genome which will likely results in low penetrance of the variants in each region of interest. It can also be due to the array used for this study (as discussed in section 3.1.3) as the design of the array means that LD blocks might not always be typed properly or by several SNPs, which might explain why we did not see groups of peaking in the Manhattan plots.

3.1.6 Parent of Origin test

The parent of origin test was performed in order to identify markers and regions of the genome that might be preferentially transmitted from mother to daughter.

The parent of origin test returned significant results for almost all the chromosomes in the the pig genome. Our significance threshold for this test was set at $p - value <= 0.003$.

- For line B, 26 SNPs reached significance on chromosome 1, 2, 3, 4, 6, 7, 8, 10, 11, 13, 14, 15 and 17, see table 3.12 for more details.
- For line C, 21 SNPs reached significance on chromosomes 1, 2, 3, 4, 6, 8, 9, 10, 13, 14, 16, 17 and 18, see table 3.13 for more details.
- For line D, 22 SNPs reached significance on chromosomes 1, 2, 3,4, 5, 6, 7, 9, 13, 15, 16 and 18 see table 3.9 for more details.
- For line H, 56 SNPs reached significance on chromosomes 1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 17 and 18 see table 3.14 for more details.

Figures 3.1, 3.12, 3.14 and 3.16 are the Manhattan plots for the parent of origin test for line B, C, D and H respectively. Figures 3.11, 3.13, 3.15 and 3.17 show the QQ plot for the p-value of the parent of origin test for each line. Table 3.15 show the lambda value linked to the QQ plot for the parent of origin test. The lambda value suggest that restricting the set to heterozygous parents increase the level of stratification present in the data given that all of the lambda values are below 0.9.

SNP ID	Chromosome	Position	P-value maternal transmission	Maternal transmitted : untransmitted counts
ASGA0001721	1	29,115,029	0.0007962	2.5:17.5
CASI0003440	1	250,635,066	0.0009111	0.5:11.5
ASGA0003663	1	92,026,588	0.001496	00:11
ALGA0113319	2	121,389,974	0.0007962	2.5:17.5
DIAS0004469	2	6,275,661	0.0009674	02:16
MARC0059849	2	125,760,041	0.002282	01:12
ALGA0018346	3	30,860,487	0.000685	1.5:15.5
MARC0009481	3	75,394,519	0.002282	01:12
ALGA0025931	4	80,915,079	0.0002896	2.5:19.5
ALGA0023973	4	21,978,564	0.001154	1.5:14.5
ALGA0035224	6	38,548,262	0.002282	01:12
INRA0026430	7	75,430,748	0.0009674	02:16
MARC0033269	8	138,870,354	0.001616	02:15
MARC0036135	10	55,403,526	0.0009674	02:16
ASGA0051058	11	66,142,455	0.0009111	02:15
ALGA0061364	11	22,730,003	0.001616	00:11
ALGA0074022	13	215,030,086	0.000532	17:02
ALGA0067709	13	7,175,532	0.0005791	02:15
H3GA0036785	13	77,651,252	0.001616	00:12
ASGA0062605	14	34,991,538	0.0009111	4.5:20.5
INRA0047046	14	129,849,565	0.001374	02:14
ALGA0075730	14	18,958,926	0.0027	00:11
ASGA0065780	14	113,376,632	0.0027	02:14
ALGA0101465	15	149,489,663	0.001496	0.5:11.5
MARC0045894	15	52,614,585	0.001616	02:15
ALGA0093735	17	23,923,753	0.001374	4.5:20.5

Table 3.12: Parent of origin results table for line B. The p-value is calculated for preferential maternal transmission against paternal transmission. The transmitted: untransmitted count are for the reference allele against the alternative allele. The 0.5 values are assigned when the parents of an individual are both heterozygotes.

SNP ID	Chromosome	Position	P-value maternal transmission	Maternal transmitted : untransmitted counts
ASGA0003310	1	76,114,707	0.00225	17.5:3.5
ASGA0009211	2	12,927,978	0.002282	01:12
ASGA0010625	2	86,139,077	0.00286	16:03
ASGA0094535	3	7,784,004	0.002282	01:12
ALGA0017723	3	14,332,984	0.0009111	11:00
H3GA0011907	4	10,423,196	0.002282	01:12
INRA0015059	4	83,749,465	0.002838	04:18
ASGA0096127	6	14,598,312	0.002183	15.5:2.5
ALGA0049156	8	117,824,360	0.00286	03:16
ALGA0056155	9	73,923,049	0.00286	03:16
MARC0082453	9	102,552,703	0.0009111	11:00
ASGA0044289	9	121,182,071	0.00286	03:16
H3GA0030271	10	52,456,152	0.00225	17.5:3.5
ALGA0070825	13	76,614,616	0.002282	01:12
MARC0066830	13	202,381,472	0.0009111	00:11
DRGA0014290	14	101,566,065	0.00225	17.5:3.5
DRGA0016061	16	32,726,222	0.001616	02:15
DRGA0016602	17	19,801,781	0.002282	01:12
ALGA0095454	17	53,299,192	0.0027	00:09
DRGA0016933	18	24,930,075	0.00225	17.5:3.5
M1GA0023271	18	50,032,308	0.0003115	00:13

Table 3.13: Parent of origin results for line C. The p-value is calculated for preferential maternal transmission. The transmitted: untransmitted count are for the reference allele against the alternative allele. The 0.5 values are assigned when the parents of an individual are both heterozygotes.

SNP ID	Chromosome	Position	P-value maternal transmission	Maternal transmitted : untransmitted counts
ALGA0002856	1	45,648,377	0.001063	18:03
ALGA0114601	1	181,402,872	0.0009111	00:11
DIAS0004495	1	245,008,421	0.0004653	01:15
H3GA0006383	2	27,576,958	4.23E-06	01:24
ASGA0104575	2	138,422,188	0.000685	15.5:1.5
ASGA0082103	3	19,116,155	0.001063	18:03
ALGA0018634	3	45,593,990	1.08E-05	1.5:23.5
ASGA0021239	4	105,898,886	0.000685	15.5:1.5
ASGA0021997	4	119,974,157	0.0027	02:14
H3GA0015082	5	1,087,963	0.0027	02:14
ASGA0026863	5	99,716,210	0.0027	02:14
MARC0009578	6	112,552,637	0.002569	0.5:10.5
H3GA0019527	7	3,813,663	0.001496	0.5:11.5
ASGA0092371	9	120,274,192	0.001616	02:15
DRGA0013265	13	190,391,508	0.0027	14:02
MARC0066830	13	202,381,472	0.000512	0.5:13.5
ALGA0074022	13	215,030,086	0.0027	00:09
MARC0101508	15	11,904,651	6.33E-05	00:16
ALGA0085056	15	46,315,392	0.0027	02:14
ASGA0073960	16	72,751,299	0.0027	02:14
ASGA0074106	16	75,024,639	0.001616	15:02
MARC0039970	18	26,878,437	0.0009111	00:11

Figure 3.9: Parent of origin results for line D. The p-value is calculated for preferential maternal transmission. The transmitted: untransmitted count are for the reference allele against the alternative allele. The 0.5 values are assigned when the parents of an individual are both heterozygotes.

SNP ID	Chromosome	Position	P-value maternal transmission	Maternal transmitted : untransmitted counts
DRGA0000439	1	33,957,887	0.002263	07:24
DRGA0000977	1	65,306,294	0.001946	1.5:13.5
ALGA0005867	1	139,478,386	5.36E-07	06:40
ALGA0114601	1	181,402,872	5.51E-09	00:34
INRA0005278	1	191,640,018	0.001154	1.5:14.5
INRA0005299	1	193,664,917	0.001154	1.5:14.5
INRA0005383	1	200,443,140	0.002497	06:22
DRGA0002016	1	250,323,029	2.73E-06	00:22
ALGA0009384	1	284,399,778	0.001946	1.5:13.5
ALGA0009414	1	284,478,293	0.001154	1.5:14.5
ALGA0017693	3	14,134,797	0.001565	10:00
H3GA0008948	3	17,859,317	0.001069	22:05
ALGA0018634	3	45,593,990	5.31E-05	03:24
MARC0053170	3	77,889,376	0.001015	06:24
DRGA0004106	3	106,777,915	0.00286	03:16
ALGA0021723	3	137,881,512	0.001154	1.5:14.5
ASGA0017039	4	2,225,565	0.00286	03:16
INRA0013709	4	38,857,169	0.001225	6.5:24.5
ALGA0025762	4	77,502,926	0.001595	06:23
ASGA0020726	4	96,803,099	0.002263	24:07
INRA0018313	5	7,507,469	0.001069	05:22
ASGA0025833	5	66,355,133	0.0027	09:00
ASGA0025896	5	67,510,148	0.0027	09:27
ALGA0034926	6	23,679,290	0.000532	00:12
H3GA0052806	6	157,353,343	4.59E-06	00:21
ASGA0034277	7	63,617,428	0.001911	6.5:23.5
M1GA0010553	7	95,143,760	2.21E-06	3.5:31.5
ASGA0037929	8	15,992,070	0.001946	1.5:13.5
ASGA0090387	8	89,832,296	0.002967	6.5:22.5
ALGA0113503	8	94,427,398	0.001946	1.5:13.5
MARC0002586	9	26,578,970	0.001225	6.5:24.5
ALGA0113506	9	33,331,814	0.001946	1.5:13.5
MARC0091645	9	128,963,159	0.001946	1.5:13.5
DRGA0009874	9	140,003,794	0.0027	09:00
ALGA0106214	9	149,004,810	0.0003182	3.5:21.5
ASGA0047525	10	36,537,382	0.001154	1.5:14.5
ASGA0047527	10	36,556,089	0.001154	1.5:14.5
H3GA0029984	10	43,738,907	2.31E-06	1.5:26.5
H3GA0033690	12	16,816,869	0.001911	6.5:23.5
H3GA0056180	12	41,548,842	0.001341	01:13
H3GA0034644	12	52,042,303	0.001069	05:22
DRGA0011849	13	978,869	0.001911	6.5:23.5
ALGA0068980	13	26,957,024	0.0007829	6.5:25.5
ALGA0069602	13	38,756,938	0.001154	1.5:14.5
MARC0020205	13	189,818,417	0.0008561	4.5:21.5
MARC0066830	13	202,381,472	7.74E-06	00:20
H3GA0040445	14	62,339,756	0.001565	10:00
ALGA0078822	14	79,562,216	0.0027	09:27
MARC0101508	15	11,904,651	5.75E-06	02:26
ALGA0084124	15	19,520,226	0.00112	7.5:26.5
ASGA0092912	15	143,902,435	0.0027	05:20
H3GA0054837	16	75,201,733	0.001946	1.5:13.5
H3GA0048171	17	29,170,040	0.001154	1.5:14.5
ALGA0095137	17	46,692,167	0.0022	4.5:19.5
ALGA0124479	18	2,082,165	0.002967	6.5:22.5
ASGA0078760	18	8,770,985	0.002282	12:01

Table 3.14: Parent of origin results for line H. The p-value is calculated for preferential maternal transmission. The transmitted: untransmitted count are for the reference allele against the alternative allele. The 0.5 values are assigned when the parents of an individual are both heterozygotes.

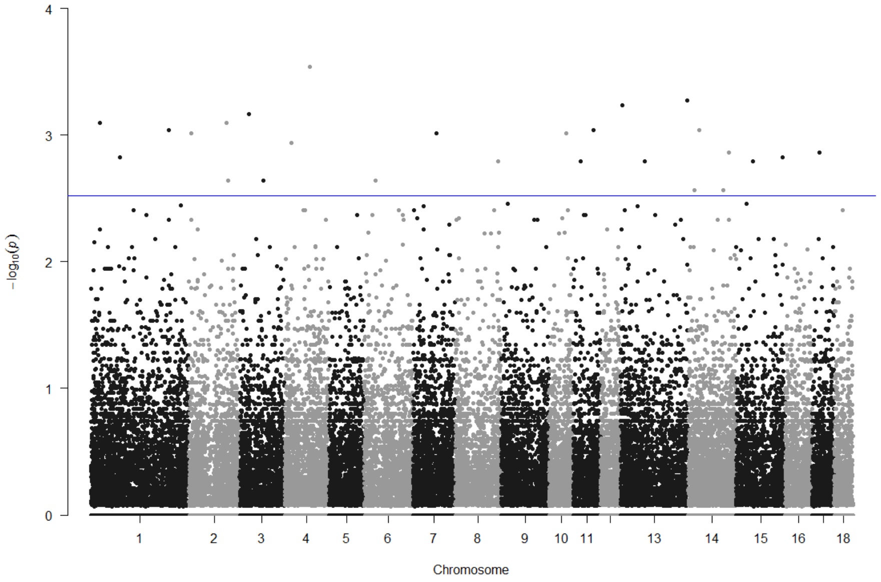


Figure 3.10: Manhattan plot for line B, parents of origin test, the blue line shows the threshold chosen to select target SNPs. The genome wide significance threshold is not displayed as it off the chart

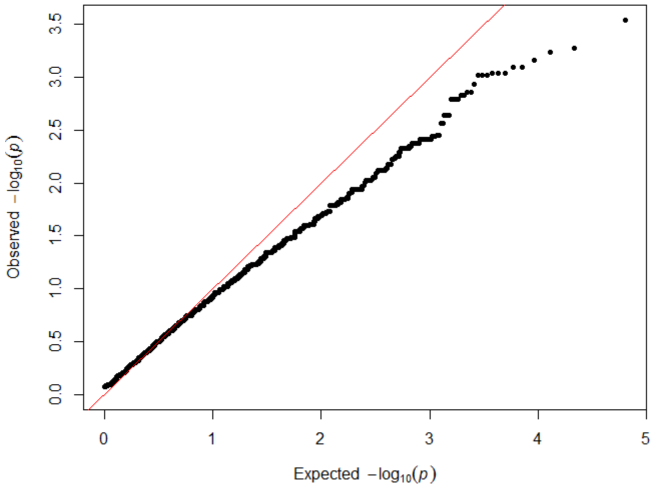


Figure 3.11: QQ plot for the p-value of the parent of origin test, for line B. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 0.879.

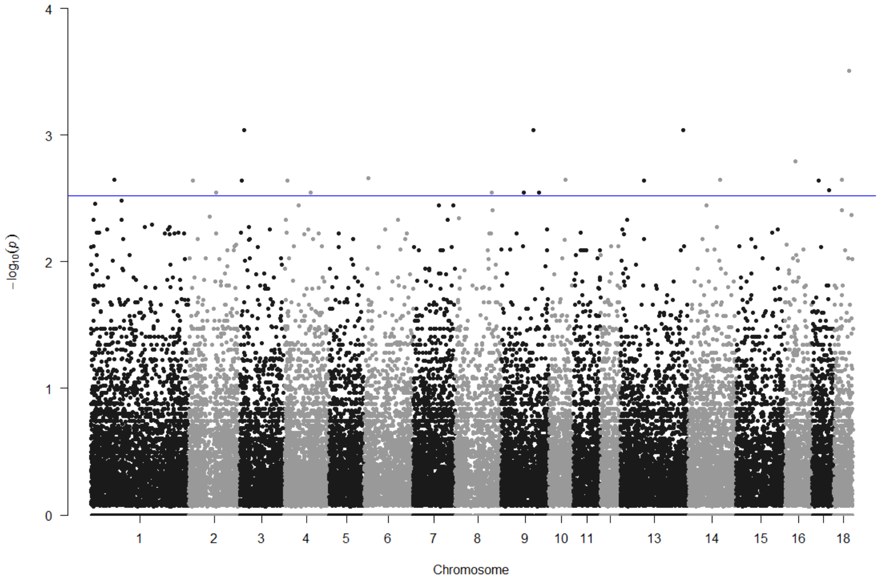


Figure 3.12: Manhattan plot for line C, parent of origin test, the blue line shows the threshold chosen to select target SNPs. The genome wide significance threshold is not displayed as it off the chart.

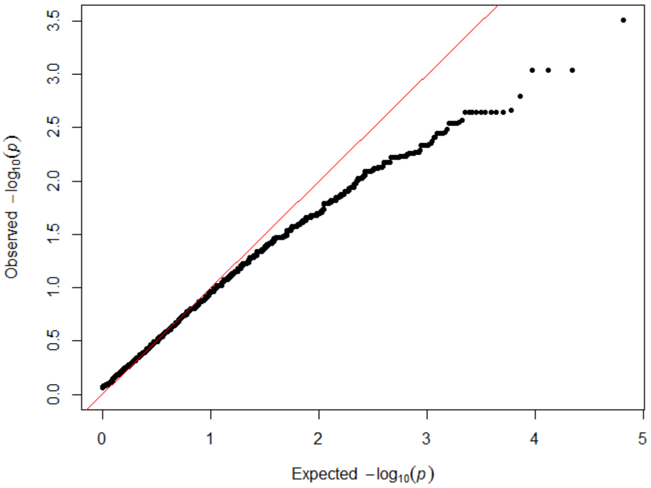


Figure 3.13: QQ plot for the p-value of the parent of origin test, for line C. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 0.879.

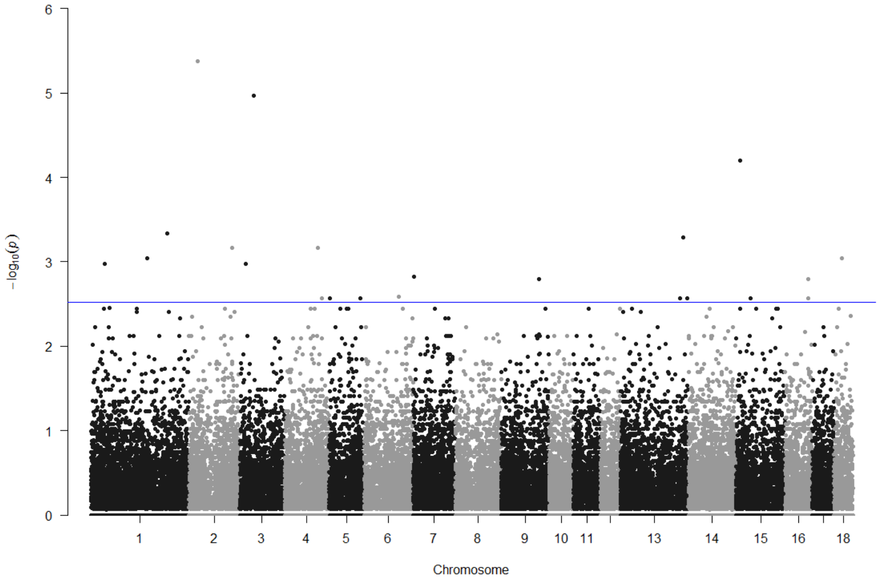


Figure 3.14: Manhattan plot for line D, parent of origin, the blue line shows the threshold chosen to select target SNPs. The genome wide significance threshold is not displayed as it off the chart.

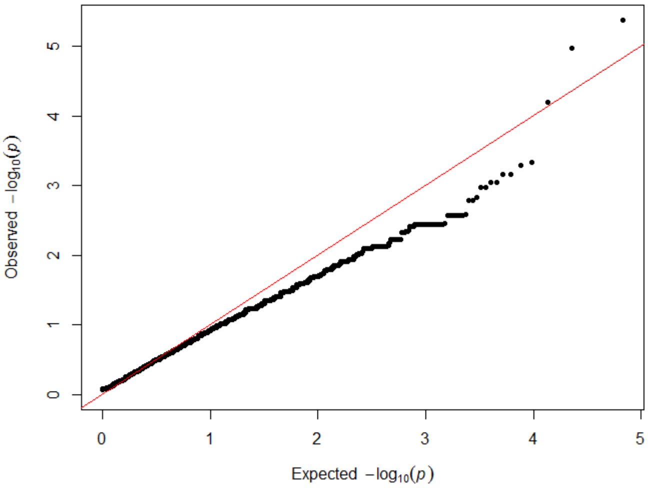


Figure 3.15: QQ plot for the p-value of the parent of origin test, for line D. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 0.8242.

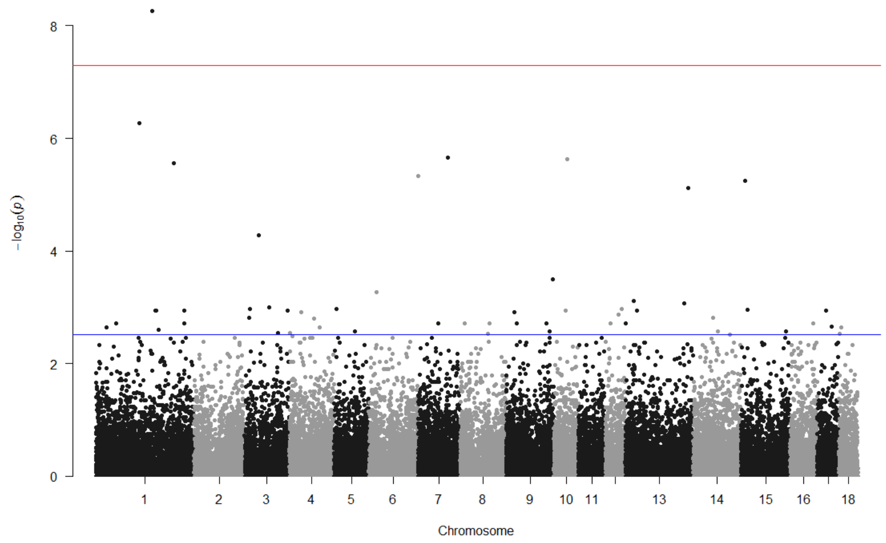


Figure 3.16: Manhattan plot for line H, parent of origin test, the blue line shows the threshold chosen to select target SNPs, the red line is genome wide significance.

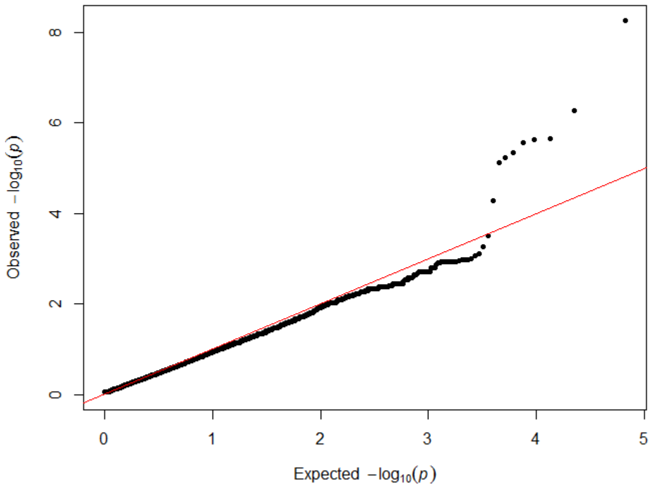


Figure 3.17: QQ plot for the p-value of the parent of origin test, for line H. Expected $-log_{10}(pvalue)$ against observed $-log_{10}(pvalue)$. The lambda value for the p-value distribution is 0.799.

Line	lambda
B	0.879
C	0.879
D	0.824
H	0.799

Table 3.15: Lambda value for the p-value distribution of the parent of origin test, all lines.

Selecting SNPs according to the thresholds defined earlier in this section and discussed in section 3.1.4, a similar number of SNPs were selected for lines B, C and D with around 20 SNPs for each.

Line H stands out with 56 SNPs selected. This might be the consequence of the higher number of families available, or rather mother-daughter pairs that were studied, as line H has the largest number of families among the lines, with 96. Our significance threshold is fairly lax: however some of the SNPs identified are common with the FBAT analysis. Similarly to the FBAT analysis, the PO analysis was designed in order to identify regions of interest for the sequence capture. The ratio of preferential transmission from mother to daughter for one allele compared to the other varies between 34:1 to 3:1.

The Manhattan plots are similar to the the FBAT ones with no clear peaks of several consecutive SNPs. As the threshold for significance is lower, the SNPs selected are closer to the background noise of non significant SNPs. Again this might be due to the low penetrance of SNPs linked to our phenotype of interest or the nature of the array. Furthermore, we were being quite lenient for the filtering of this results as the regions and genes identified will be used for the design of the capture set which will allow us to confirm or exclude the candidate regions previously identified.

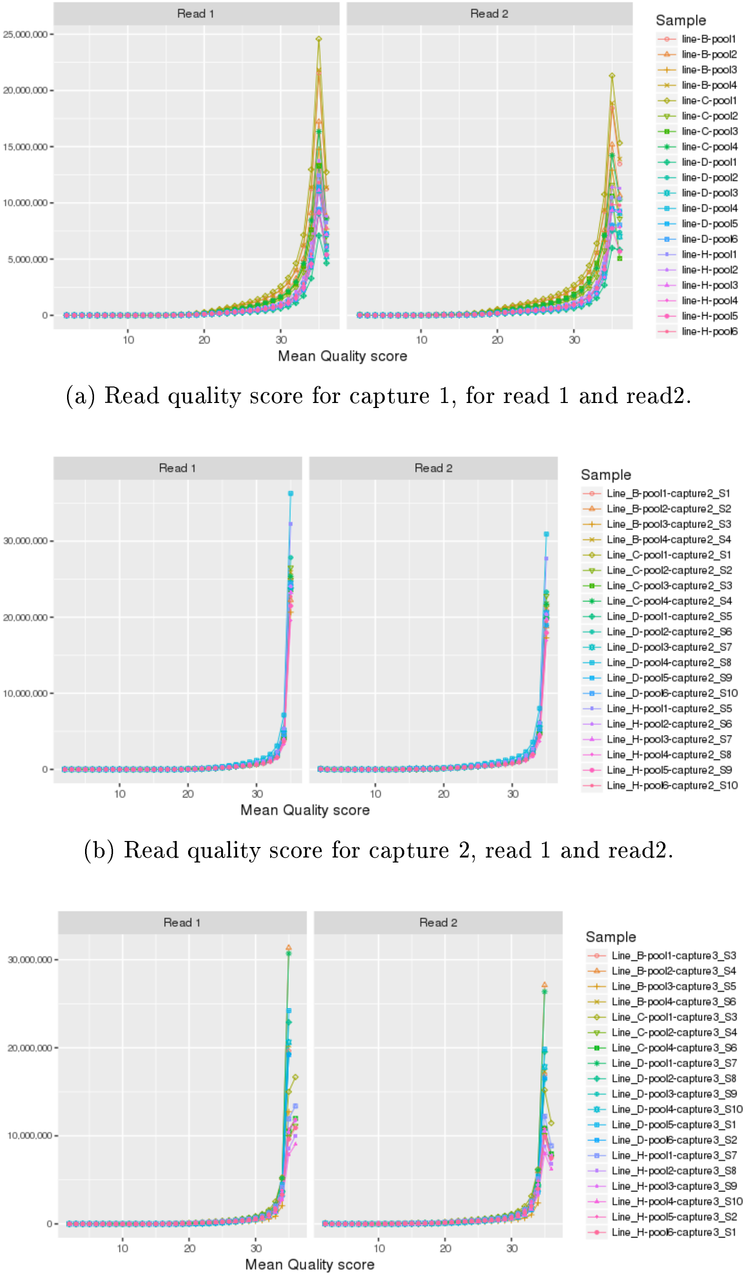
3.2 Sequencing data QC

3.2.1 Run and mapping metrics

As discussed in material and methods, the animals were divided between different pools. For line B and C the first pool is composed of animals with a history of infanticide sows in their family, pool 2 for sows who have several infanticide episodes and pool 3 and 4 for control samples. Lines D and H have 3 pools for animals with a history of infanticide (pool 1, 2 and 3), pool 4 is for the serial infanticidle and pool 5 and 6 for controls.

3.2.1.1 Read quality

The mean read quality scores for the pool in each of the capture sets are displayed in figure 3.18. The quality is generally above a Qscore of 30 for all the samples. A Qscore of 30 represents a 1 in a 1000 chance of having a wrong base call. The capture 1 set has slightly lower quality reading than the other two capture sets. The per base sequence quality score is given in figures 3.19a and 3.19b for capture set 1line B, the rest of the plots are available in the supplementary material. All of the samples share a similar pattern for these plots.

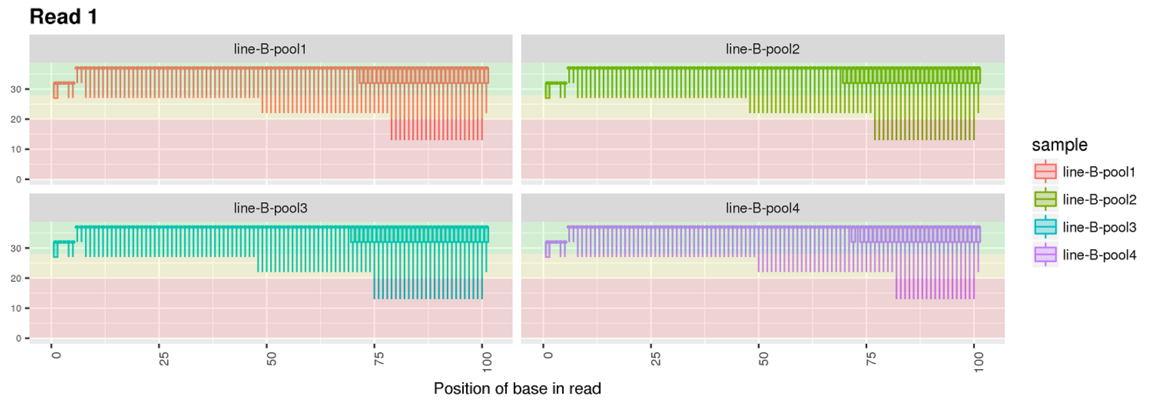


(a) Read quality score for capture 1, for read 1 and read2.

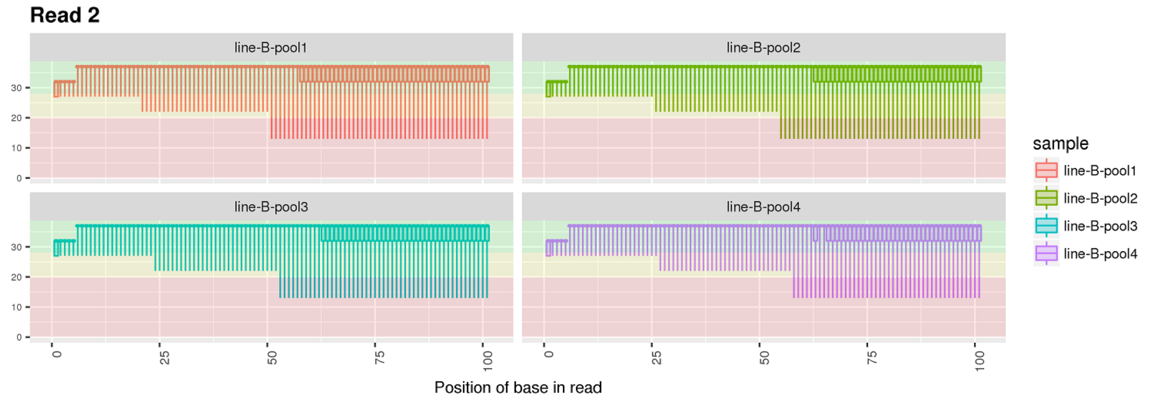
(b) Read quality score for capture 2, read 1 and read2.

(c) Read quality score for capture 3, read 1 and read 2.

Figure 3.18: Read quality score for the capture set. Figure 3.18a for capture 1, Figure 3.18b for capture 2 and Figure 3.18c for capture 3



(a) Read 1 quality score per based, capture set 1 line B



(b) Read 2 quality score per base, capture set 1 line B

Figure 3.19: Read quality score per base, read 1 and read 2 of capture set 1 line B

3.2.1.2 Mapping QC

The number of reads for each of the capture and pool is summarized in figures 3.20, 3.21 and 3.22 for capture set 1, 2 and 3 respectively. The pools are the same as presented before. The same figures for read 1 and read 2 individually are available in the supplementary material. The capture set 1 is the one showing the largest amount of variation for the number of reads mapped to the genome. It varies between 30 million to 150 million per samples.

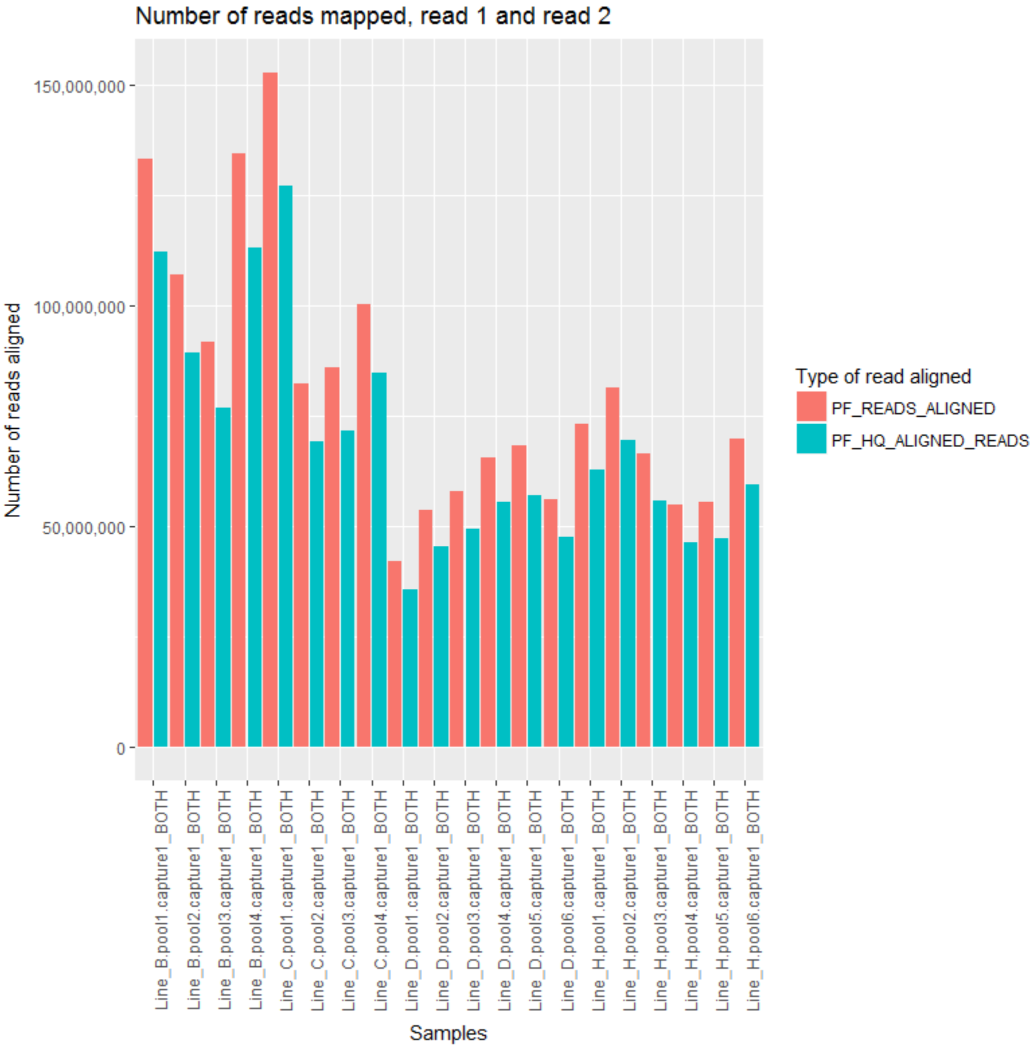


Figure 3.20: Number of reads mapped from both ends (Read 1 and Read 2) for Capture set 1; pools 1, 2, 3 and lines B, C, D and H. Two sets of data are displayed for read passing filter (PF, passing Illumina chastity filter, see section 4.2.3.1). In red are all the reads PF that aligned and in blue are the high quality (HQ) reads PF that aligned. High quality is defined by a mapped quality score of 20 (Q20) or above . This represents 1/100 or smaller chance of the alignment being wrong.

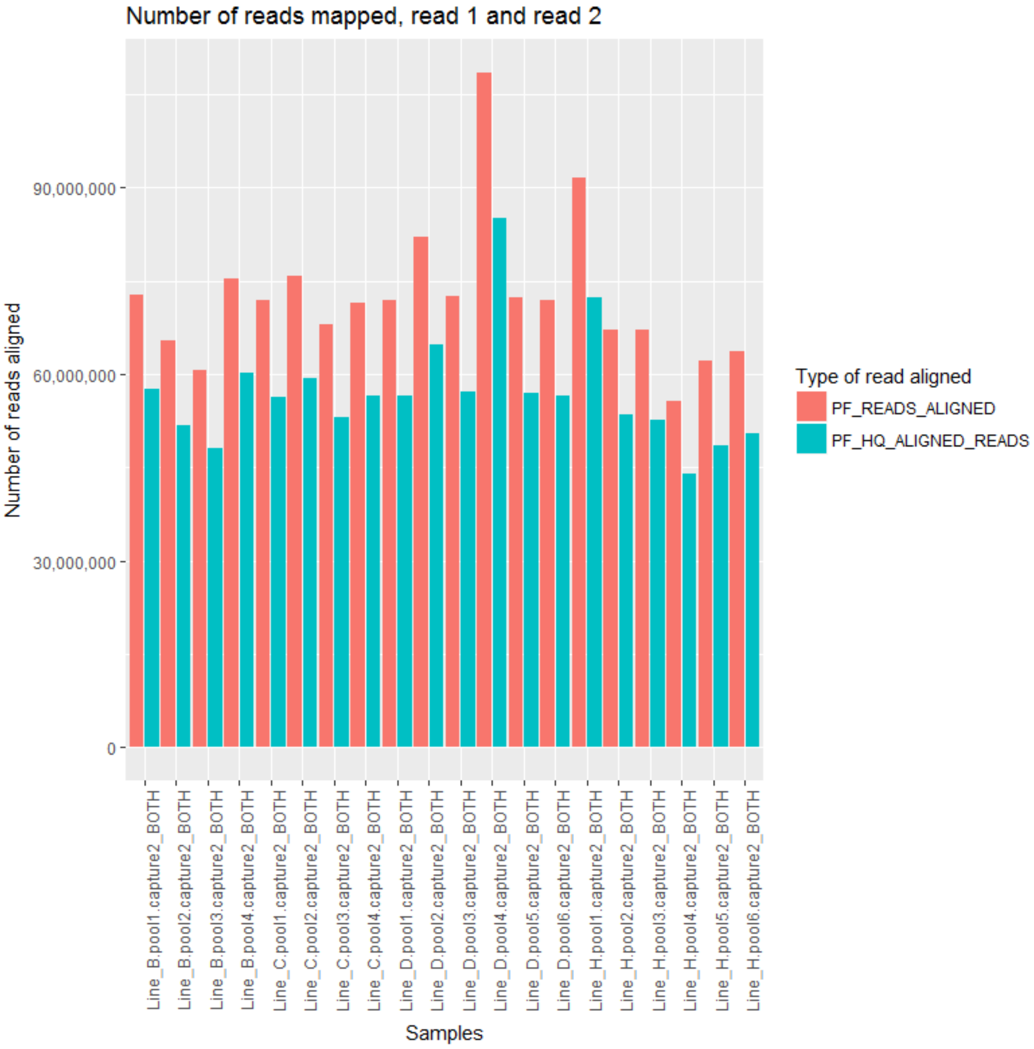


Figure 3.21: Number of reads mapped from both ends (Read 1 and Read 2) for Capture set 2; pools 1, 2, 3 and lines B, C, D and H. Two sets of data are displayed for read passing filter (PF, passing Illumina chastity filter, see section 4.2.3.1). In red are all the reads PF that aligned and in blue are the high quality (HQ) reads PF that aligned. High quality is defined by a mapped quality score of 20 (Q20) or above . This represents 1/100 or smaller chance of the alignment being wrong.

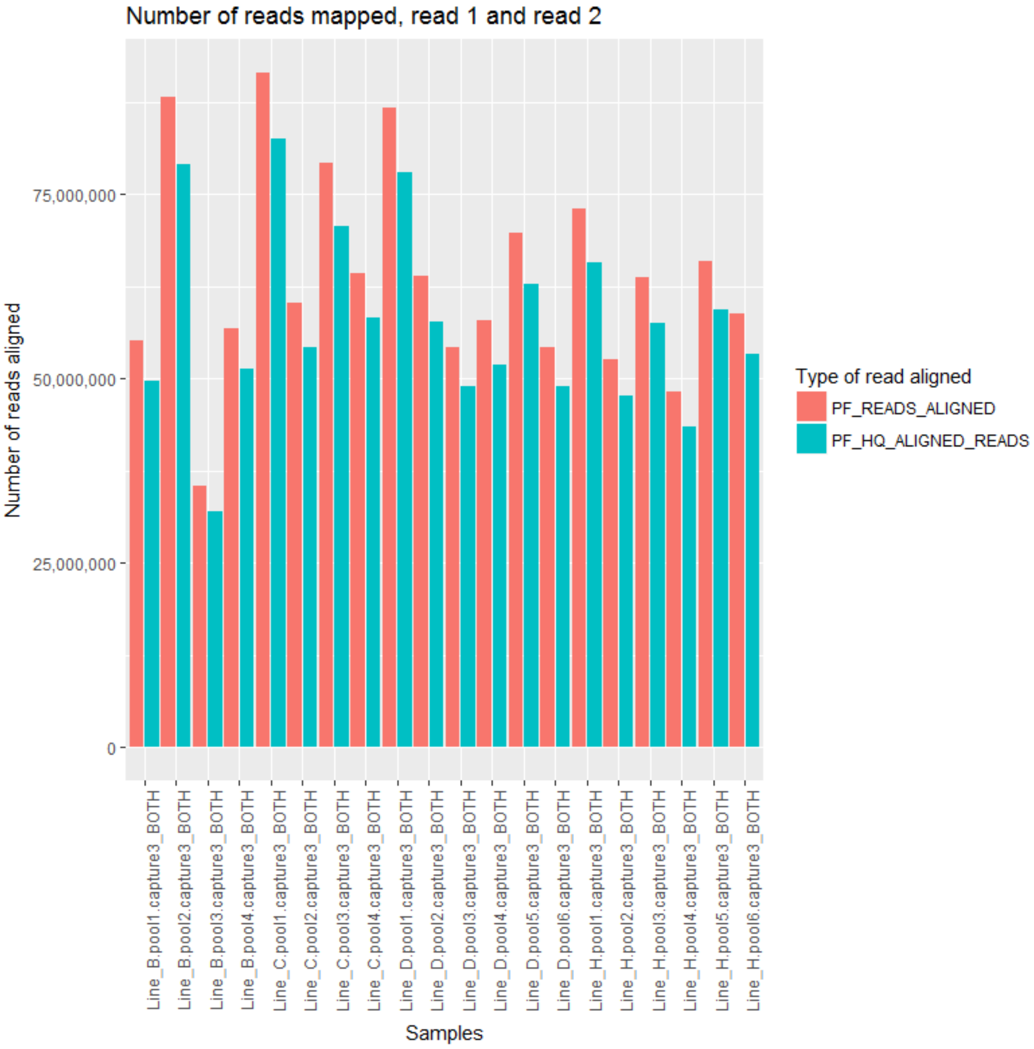


Figure 3.22: Number of reads mapped from both ends (Read 1 and Read 2) for Capture set 3; pools 1, 2 and 3 and lines B, C, D and H. Two sets of data are displayed for read passing filter (PF, passing Illumina chastity filter, see section 4.2.3.1). In red are all the reads PF that aligned and in blue are the high quality (HQ) reads PF that aligned. High quality is defined by a mapped quality score of 20 (Q20) or above . This represents 1/100 or smaller chance of the alignment being wrong.

Tables 3.16, 3.17 and 3.18 give the summary of the number of reads mapped for each capture set, combining read 1 and read 2. Passing filter high quality aligned reads are reads aligning with an quality score (Qscore) above 20.

- Capture set 1, average number of reads passing filter mapped: 81,681,252, minimum 42,266,915 reads and maximum 152,667,192. Average percentage read aligned is 97% . For high quality read the average number of aligned reads is 68,890,382 reads, minimum 35,786,637 reads and maximum 127,159,614 reads.

- Capture set 2, average number of reads passing filter mapped is 72,457,794, the minimum number of reads mapped is 55,729,390 and the maximum 108,435,627. The average percentage of aligned read is 98%. High quality reads the average number of aligned reads is 57,168,410 with a minimum of 43,923,060 and a maximum of 85,187,732 reads.

- Capture set 3, average numbe of reads passing filger mapped is 64,030,227, the minimum number of reads mappis is 35,543,830 and the maximum 91,433,220. The average percentatage of read aligned is 98.1%. High quality reads the average number of aligned reads is 57,664,615 with a minimum of 32,054,462 and a maximum of 82,476,830 reads.

Lines	TOTAL_READS	PF_READS aligned	PCT_PF_READS aligned	PF_HQ_READS aligned	PF MISMATCH_RATE
Line H-pool1-capture1	75,088,750	73,327,924	0.9766	62,997,796	0.0101
Line H-pool2-capture1	83,379,614	81,383,067	0.9761	69,429,083	0.0103
Line H-pool3-capture1	67,990,646	66,452,833	0.9774	55,922,274	0.0106
Line H-pool4-capture1	56,327,248	55,023,992	0.9769	46,530,743	0.0105
Line H-pool5-capture1	57,058,626	55,578,500	0.9741	47,240,597	0.0113
Line H-pool6-capture1	71,676,552	69,898,000	0.9752	59,427,771	0.0104
Line B-pool1-capture1	137,314,694	133,121,532	0.9695	112,316,531	0.0127
Line B-pool2-capture1	110,235,764	106,933,563	0.97	89,472,228	0.0128
Line B-pool3-capture1	94,566,318	91,792,575	0.9707	76,896,859	0.0128
Line B-pool4-capture1	138,370,378	134,476,470	0.9719	113,246,554	0.0126
Line C-pool1-capture1	157,301,982	152,667,192	0.9705	127,159,614	0.013
Line C-pool2-capture1	84,948,514	82,485,978	0.971	69,202,084	0.0126
Line C-pool3-capture1	88,607,772	86,006,584	0.9706	71,825,218	0.0141
Line C-pool4-capture1	103,359,932	100,497,895	0.9723	84,713,674	0.0125
Line D-pool1-capture1	43,378,990	42,266,915	0.9744	35,786,637	0.0107
Line D-pool2-capture1	54,861,480	53,609,029	0.9772	45,524,947	0.0109
Line D-pool3-capture1	59,594,480	58,121,745	0.9753	49,595,063	0.0119
Line D-pool4-capture1	67,319,302	65,585,312	0.9742	55,673,911	0.0108
Line D-pool5-capture1	70,140,066	68,279,765	0.9735	57,156,781	0.0113
Line D-pool6-capture1	57,437,276	56,116,171	0.977	47,689,276	0.0107

Table 3.16: Capture 1, read 1 and read 2 combined, mapping statistics. TOTAL_READS is the total number of reads generated passing filter. PF_READS aligned the number of reads passing filter aligned. PCT_PF_READ aligned is the percentage of reads passing filter aligned. PF_HQ_READS aligned is the number of reads passing filter aligned and considered to be of high quality. PF MISMATCH_RATE is the percentage of mismatches in the reads passing filter.

Lines	TOTAL_READS	PF_READS aligned	PCT_PF_READS aligned	PF_HQ_READS aligned	PF MISMATCH_RATE
Line H-pool1-capture2	93,025,096	91,594,335	0.9846	72,378,498	0.0061
Line H-pool2-capture2	68,337,872	67,095,960	0.9818	53,439,245	0.0062
Line H-pool3-capture2	68,414,060	67,236,953	0.9828	52,651,322	0.0065
Line H-pool4-capture2	56,878,630	55,729,390	0.9798	43,923,060	0.0065
Line H-pool5-capture2	63,228,748	62,122,086	0.9825	48,495,196	0.0065
Line H-pool6-capture2	64,582,542	63,649,267	0.9855	50,525,936	0.0060
Line B-pool1-capture2	74,177,346	72,740,408	0.9806	57,747,087	0.0065
Line B-pool2-capture2	66,656,046	65,357,046	0.9805	51,877,093	0.0065
Line B-pool3-capture2	61,852,196	60,692,535	0.9813	48,225,325	0.0064
Line B-pool4-capture2	76,755,578	75,456,834	0.9831	60,347,383	0.0063
Line C-pool1-capture2	73,501,254	71,824,670	0.9772	56,412,312	0.0073
Line C-pool2-capture2	77,280,472	75,858,941	0.9816	59,401,054	0.0070
Line C-pool3-capture2	69,341,074	68,140,278	0.9827	53,200,939	0.0069
Line C-pool4-capture2	72,893,000	71,596,541	0.9822	56,590,495	0.0067
Line D-pool1-capture2	73,722,136	71,966,577	0.9762	56,578,548	0.0068
Line D-pool2-capture2	83,885,058	82,091,472	0.9786	64,885,960	0.0065
Line D-pool3-capture2	74,011,940	72,651,369	0.9816	57,164,190	0.0064
Line D-pool4-capture2	110,664,764	108,435,627	0.9799	85,187,732	0.0065
Line D-pool5-capture2	74,257,008	72,271,579	0.9733	56,995,236	0.0067
Line D-pool6-capture2	73,238,120	71,838,954	0.9809	56,661,413	0.0063

Table 3.17: Capture 2, read 1 and 2 combined, mapping statistics. TOTAL_READS is the total number of reads generated passing filter. PF_READS aligned the number of reads passing filter aligned. PCT_PF_READ aligned is the percentage of reads passing filter aligned. PF_HQ_READS aligned is the number of reads passing filter aligned and considered to be of high quality. PF MISMATCH_RATE is the percentage of mismatches in the reads passing filter.

Lines	TOTAL_READS	PF_READS aligned	PCT_PF_READS aligned	PF_HQ_READS aligned	PF MISMATCH_RATE
Line H-pool1-capture3	74,867,550	73,104,287	0.9764	65,732,765	0.0085
Line H-pool2-capture3	53,778,638	52,656,091	0.9791	47,746,163	0.0083
Line H-pool3-capture3	65,054,464	63,736,784	0.9797	57,489,322	0.0084
Line H-pool4-capture3	49,278,188	48,259,756	0.9793	43,451,734	0.0084
Line H-pool5-capture3	67,599,424	66,019,235	0.9766	59,321,419	0.0088
Line H-pool6-capture3	60,046,120	58,845,062	0.9800	53,420,134	0.0083
Line B-pool1-capture3	55,924,538	55,096,376	0.9852	49,785,634	0.0063
Line B-pool2-capture3	89,544,320	88,146,755	0.9844	79,035,370	0.0065
Line B-pool3-capture3	36,073,302	35,543,830	0.9853	32,054,462	0.0064
Line B-pool4-capture3	57,614,252	56,832,081	0.9864	51,406,889	0.0063
Line C-pool1-capture3	93,458,664	91,433,220	0.9783	82,476,830	0.0090
Line C-pool2-capture3	61,589,512	60,327,979	0.9795	54,284,495	0.0089
Line C-pool3-capture3	81,305,174	79,342,297	0.9759	70,643,145	0.0092
Line C-pool4-capture3	65,680,586	64,362,304	0.9799	58,239,392	0.0087
Line D-pool1-capture3	88,065,810	86,746,177	0.9850	78,063,598	0.0059
Line D-pool2-capture3	64,940,980	63,913,754	0.9842	57,643,778	0.0060
Line D-pool3-capture3	55,108,940	54,210,065	0.9837	48,966,302	0.0060
Line D-pool4-capture3	58,824,874	57,916,962	0.9846	51,847,229	0.0060
Line D-pool5-capture3	71,074,430	69,803,072	0.9821	62,796,187	0.0063
Line D-pool6-capture3	55,273,464	54,308,452	0.9825	48,887,455	0.0062

Table 3.18: Capture 3, read 1 and 2 combined mapping statistics. TOTAL_READS is the total number of reads generated passing filter. PF_READS aligned the number of reads passing filter aligned. PCT_PF_READ aligned is the percentage of reads passing filter aligned. PF_HQ_READS aligned is the number of reads passing filter aligned and considered to be of high quality. PF MISMATCH_RATE is the percentage of mismatches in the reads passing filter.

3.2.1.3 Sequence duplication levels.

The level of duplication for the 3 capture sets is plotted in figures 3.23, 3.24 and 3.25 for capture sets 1, 2 and 3 respectively. The levels of duplications are fairly similar between samples, with a peak for five to ten percent of the reads being duplicated between 10 and 50x. Tables 3.19, 3.20, 3.21 show the percentage of reads left after deduplication, the total number of reads and the estimated number of read after deduplication. This estimate is based on a sample of 100,000 reads. The percentage of reads left is between 40% and 75% and the spread of reads in the pools after demultiplexing is estimated at between 11 millions to 48 millions reads.

Line and pool	Read	Percentage left after deduplication	Total number of read	Read left after deduplication
line-B-pool1	R1	58.79	68,657,347	40,363,654
line-B-pool1	R2	60.46	68,657,347	41,510,232
line-B-pool2	R1	62.46	55,117,882	34,426,629
line-B-pool2	R2	63.68	55,117,882	35,099,067
line-B-pool3	R1	64.64	47,283,159	30,563,834
line-B-pool3	R2	65.69	47,283,159	31,060,307
line-B-pool4	R1	61.5	69,185,189	42,548,891
line-B-pool4	R2	62.71	69,185,189	43,386,032
line-C-pool1	R1	59.85	78,650,991	47,072,618
line-C-pool1	R2	61.69	78,650,991	48,519,796
line-C-pool2	R1	66.7	42,474,257	28,330,329
line-C-pool2	R2	67.49	42,474,257	28,665,876
line-C-pool3	R1	68.13	44,303,886	30,184,238
line-C-pool3	R2	69.96	44,303,886	30,994,999
line-C-pool4	R1	64.69	51,679,966	33,431,770
line-C-pool4	R2	65.83	51,679,966	34,020,922
line-D-pool1	R1	73.18	21,689,495	15,872,372
line-D-pool1	R2	73.36	21,689,495	15,911,414
line-D-pool2	R1	69.99	27,430,740	19,198,775
line-D-pool2	R2	70.31	27,430,740	19,286,553
line-D-pool3	R1	69.45	29,797,240	20,694,183
line-D-pool3	R2	69.87	29,797,240	20,819,332
line-D-pool4	R1	66.85	33,659,651	22,501,477
line-D-pool4	R2	66.97	33,659,651	22,541,868
line-D-pool5	R1	67.48	35,070,033	23,665,258
line-D-pool5	R2	67.9	35,070,033	23,812,552
line-D-pool6	R1	68.67	28,718,638	19,721,089
line-D-pool6	R2	68.63	28,718,638	19,709,601
line-H-pool1	R1	62.96	37,544,375	23,637,939
line-H-pool1	R2	62.96	37,544,375	23,637,939
line-H-pool2	R1	63.36	41,689,807	26,414,662
line-H-pool2	R2	63.48	41,689,807	26,464,689
line-H-pool3	R1	67.61	33,995,323	22,984,238
line-H-pool3	R2	67.93	33,995,323	23,093,023
line-H-pool4	R1	69.24	28,163,624	19,500,493
line-H-pool4	R2	69.07	28,163,624	19,452,615
line-H-pool5	R1	68.04	28,529,313	19,411,345
line-H-pool5	R2	68.35	28,529,313	19,499,785
line-H-pool6	R1	67.02	35,838,276	24,018,813
line-H-pool6	R2	67.39	35,838,276	24,151,414

Table 3.19: Capture 1 duplication estimation table. The value are calculated based on a subset of a 100,000 reads by fastQC

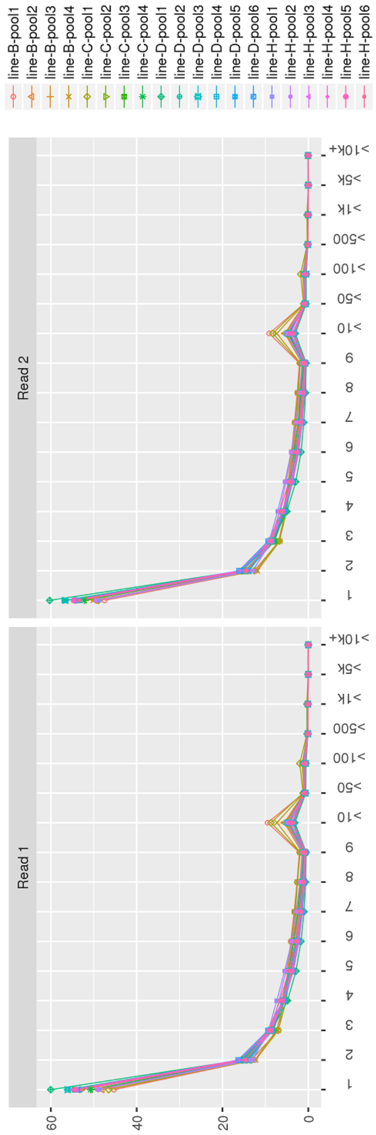


Figure 3.23: Duplication levels for capture 1 all lines, all pools.

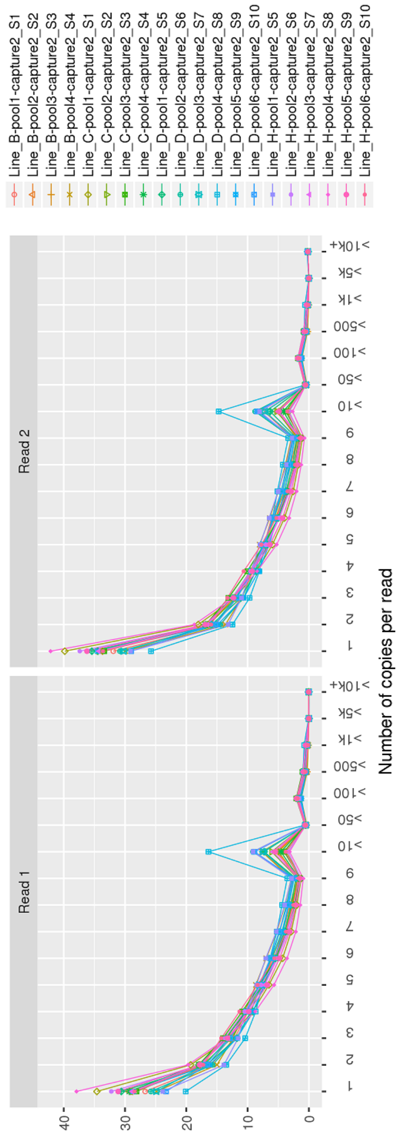


Figure 3.24: Duplication levels for capture set 2, all lines, all pools.

Line and pool	Read	Percentage left after deduplication	Total number of read	Read left after deduplication
Line_B-pool1	R1	45.91	37,088,673	17,027,410
Line_B-pool1	R2	49.49	37,088,673	18,355,184
Line_B-pool2	R1	48.72	33,328,023	16,237,413
Line_B-pool2	R2	51.92	33,328,023	17,303,910
Line_B-pool3	R1	48.89	30,926,098	15,119,769
Line_B-pool3	R2	52.21	30,926,098	16,146,516
Line_B-pool4	R1	44.18	38,377,789	16,955,307
Line_B-pool4	R2	47.59	38,377,789	18,263,990
Line_C-pool1	R1	53.97	36,750,627	19,834,313
Line_C-pool1	R2	57.8	36,750,627	21,241,862
Line_C-pool2	R1	48.1	38,640,236	18,585,954
Line_C-pool2	R2	52.33	38,640,236	20,220,435
Line_C-pool3	R1	48.25	34,670,537	16,728,534
Line_C-pool3	R2	52.15	34,670,537	18,080,685
Line_C-pool4	R1	49.2	36,446,500	17,931,678
Line_C-pool4	R2	52.95	36,446,500	19,298,422
Line_D-pool1	R1	49.2	36,861,068	18,135,645
Line_D-pool1	R2	52.94	36,861,068	19,514,249
Line_D-pool2	R1	44.67	41,942,529	18,735,728
Line_D-pool2	R2	48.3	41,942,529	20,258,242
Line_D-pool3	R1	44.8	37,005,970	16,578,675
Line_D-pool3	R2	48.6	37,005,970	17,984,901
Line_D-pool4	R1	37.93	55,332,382	20,987,572
Line_D-pool4	R2	42.39	55,332,382	23,455,397
Line_D-pool5	R1	47.32	37,128,504	17,569,208
Line_D-pool5	R2	51.5	37,128,504	19,121,180
Line_D-pool6	R1	43.2	36,619,060	15,819,434
Line_D-pool6	R2	47.31	36,619,060	17,324,477
Line_H-pool1	R1	42.38	46,512,548	19,712,018
Line_H-pool1	R2	46.49	46,512,548	21,623,684
Line_H-pool2	R1	51.76	34,168,936	17,685,841
Line_H-pool2	R2	55.47	34,168,936	18,953,509
Line_H-pool3	R1	48.4	34,207,030	16,556,203
Line_H-pool3	R2	52.5	34,207,030	17,958,691
Line_H-pool4	R1	57	28,439,315	16,210,410
Line_H-pool4	R2	60.07	28,439,315	17,083,497
Line_H-pool5	R1	50.37	31,614,374	15,924,160
Line_H-pool5	R2	54.22	31,614,374	17,141,314
Line_H-pool6	R1	48.59	32,291,271	15,690,329
Line_H-pool6	R2	52.33	32,291,271	16,898,022

Table 3.20: Capture 2 duplication table. The value are calculated based on a subset of a 100,000 reads by fastQC

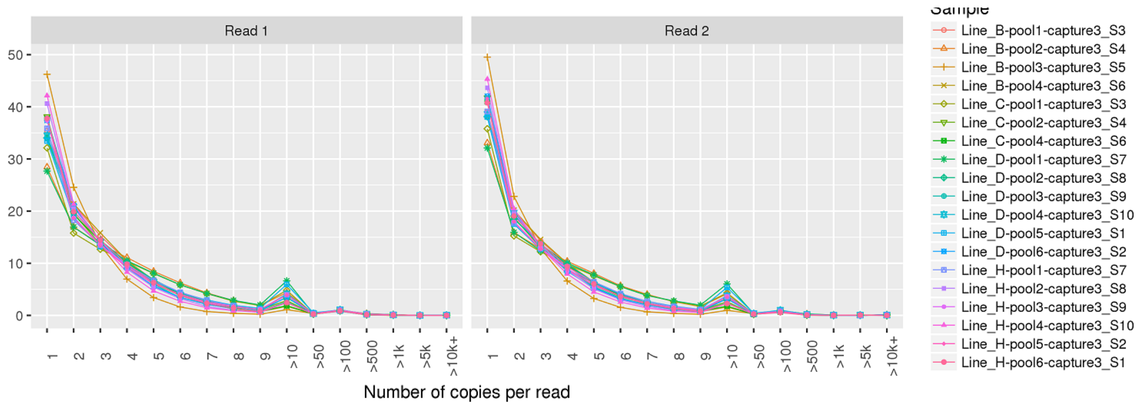


Figure 3.25: Duplication levels for capture set 3, all lines, all pools.

Line and pool	Read	Percentage left after deduplication	Total number of read	Read left after deduplication
Line_B-pool1	R1	56.33	27,962,269	15,751,146
Line_B-pool1	R2	58.83	27,962,269	16,450,203
Line_B-pool2	R1	48.49	44,772,160	21,710,020
Line_B-pool2	R2	51.85	44,772,160	23,214,365
Line_B-pool3	R1	65.95	18,036,651	11,895,171
Line_B-pool3	R2	68.01	18,036,651	12,266,726
Line_B-pool4	R1	55.51	28,807,126	15,990,836
Line_B-pool4	R2	58.1	28,807,126	16,736,940
Line_C-pool1	R1	50.96	46,729,332	23,813,268
Line_C-pool1	R2	53.81	46,729,332	25,145,054
Line_C-pool2	R1	57.51	30,794,756	17,710,064
Line_C-pool2	R2	60.2	30,794,756	18,538,443
Line_C-pool3	R1	54.62	40,652,587	22,204,443
Line_C-pool3	R2	58.03	40,652,587	23,590,696
Line_C-pool4	R1	57.47	32,840,293	18,873,316
Line_C-pool4	R2	59.8	32,840,293	19,638,495
Line_D-pool1	R1	47.54	44,032,905	20,933,243
Line_D-pool1	R2	50.75	44,032,905	22,346,699
Line_D-pool2	R1	54.4	32,470,490	17,663,947
Line_D-pool2	R2	57.2	32,470,490	18,573,120
Line_D-pool3	R1	57.91	27,554,470	15,956,794
Line_D-pool3	R2	60.5	27,554,470	16,670,454
Line_D-pool4	R1	54.79	29,412,437	16,115,074
Line_D-pool4	R2	57.71	29,412,437	16,973,917
Line_D-pool5	R1	52.94	35,537,215	18,813,402
Line_D-pool5	R2	56.26	35,537,215	19,993,237
Line_D-pool6	R1	57.43	27,636,732	15,871,775
Line_D-pool6	R2	60.56	27,636,732	16,736,805
Line_H-pool1	R1	55.05	37,433,775	20,607,293
Line_H-pool1	R2	57.54	37,433,775	21,539,394
Line_H-pool2	R1	60.11	26,889,319	16,163,170
Line_H-pool2	R2	62.23	26,889,319	16,733,223
Line_H-pool3	R1	56.45	32,527,232	18,361,622
Line_H-pool3	R2	59.19	32,527,232	19,252,869
Line_H-pool4	R1	61.49	24,639,094	15,150,579
Line_H-pool4	R2	63.71	24,639,094	15,697,567
Line_H-pool5	R1	56.72	33,799,712	19,171,197
Line_H-pool5	R2	59.61	33,799,712	20,148,008
Line_H-pool6	R1	57.61	30,023,060	17,296,285
Line_H-pool6	R2	59.78	30,023,060	17,947,785

Table 3.21: Capture 3 duplication table. The value are calculated based on a subset of a 100,000 reads by fastQC

3.2.2 Capture quality control

Summary metrics for the each capture set after deduplication are displayed in tables 3.22, 3.23, 3.24 for the different pools of sequenced animals in capture sets 1, 2 and 3. The metrics in the table are fr, ft, enrichment, average coverage, sd coverage and the 25th, 50th, 75th and 100th percentile. fr represent the fraction of reads mapping to the targetted area (capture set target). For the first capture set the average is 0.5 so around 50% of the reads map to the capture area, it increases to 64% (0.64) of reads for capture 2 and 66% (0.66) for capture set 3. For "off the shelf" capture kits, it is expected to get between 60% to 80% capture efficiency [183], however our sequence capture is a custom design and based on a fragmented genome build as described in section 1.7, the off target effect might be greater. For a more detailed discussion about this, see section 4.2.4. ft represents the fraction of reads mapping to the rest of the genome. This is really small, less than one percent for all capture sets but it is expected due to the capture strategy used. The enrichment value is calculated as the ratio between fr/ft. The enrichment values are on average 368, 387 and 365 for the pools in capture set 1, 2 and 3 respectively. The next value is the average coverage for the targeted bases in each pool. The average coverage for the pools for each capture set are 900, 874 and 748 for set 1, 2 and 3. The corresponding standard deviation averages are 457, 394 and 266. The percentile data can be used to identify any skew in the data. For all the capture sets the median (50th percentile) is relatively close to the average coverage. The 100th percentile shows the maximum coverage for the pool and some of the values are very high compared to the mean (around 10 times more). The figure suggests that the capture set performed relatively well, at least half of the reads mapping to targetted region. The coverage of the the targeted region is very high with at least 748 reads on average covering the targetted region. The 25thpercentile is still relatively high, with an average of 584, 595 and 539, suggesting that the lower range of the coverage is still high.

Sample Name	fr	ft	enrichment	average coverage	sd coverage	25%	50%	75%	100%
Line H-pool1-capture1	0.5792	0.0014	429	953	461	629	945	1257	10934
Line H-pool2-capture1	0.5390	0.0014	399	983	494	633	966	1300	10251
Line H-pool3-capture1	0.4785	0.0014	354	715	361	465	699	938	11762
Line H-pool4-capture1	0.4990	0.0014	370	616	305	409	607	804	11237
Line H-pool5-capture1	0.5458	0.0014	404	682	353	440	662	892	9052
Line H-pool6-capture1	0.5134	0.0014	380	803	381	534	799	1057	9734
Line B-pool1-capture1	0.5059	0.0014	375	1511	786	972	1477	1984	21226
Line B-pool2-capture1	0.4722	0.0014	350	1138	609	728	1109	1490	19988
Line B-pool3-capture1	0.4583	0.0014	339	945	478	619	929	1237	20574
Line B-pool4-capture1	0.4790	0.0014	355	1446	717	952	1434	1904	25132
Line C-pool1-capture1	0.4488	0.0014	332	1544	804	993	1517	2035	26313
Line C-pool2-capture1	0.4555	0.0014	337	845	427	546	832	1116	14437
Line C-pool3-capture1	0.4383	0.0014	325	849	436	550	830	1112	14579
Line C-pool4-capture1	0.4615	0.0014	342	1040	511	681	1030	1371	15217
Line D-pool1-capture1	0.5083	0.0014	377	482	249	314	470	630	8623
Line D-pool2-capture1	0.5092	0.0014	377	613	321	397	600	803	10640
Line D-pool3-capture1	0.5452	0.0014	404	711	355	467	699	932	11520
Line D-pool4-capture1	0.5233	0.0014	388	772	405	495	745	1011	13326
Line D-pool5-capture1	0.4513	0.0014	334	694	378	437	667	908	14922
Line D-pool6-capture1	0.5150	0.0014	381	649	311	437	643	848	11001

Table 3.22: Summary of capture statistics for capture set 1 and the different pools for each line. The statistics reported are fr: fraction of reads covering the target and ft fraction of reads covering the genome. Enrichiment is the ratio: fr/ft. Coverage metrics are: *average coverage:* average coverage for the target bases and *sd coverage*: coverage standard deviation. The last 4 columns display the 25th, 50th (median) , 75th and 100th percentile for coverage.

Sample Name	fr	ft	enrichment	average coverage	sd coverage	25%	50%	75%	100%
Line H-pool1-capture2	0.6665	0.0017	403	1147	603	774	1144	1511	84120
Line H-pool2-capture2	0.6310	0.0017	381	794	365	526	790	1054	6945
Line H-pool3-capture2	0.6287	0.0017	380	795	336	557	795	1027	5605
Line H-pool4-capture2	0.5906	0.0017	357	620	268	432	616	800	11011
Line H-pool5-capture2	0.6380	0.0017	385	746	330	520	740	959	13053
Line H-pool6-capture2	0.6873	0.0017	415	821	394	535	811	1092	19751
Line B-pool1-capture2	0.6643	0.0017	401	909	432	595	898	1210	6319
Line B-pool2-capture2	0.6715	0.0017	406	826	373	558	822	1088	5199
Line B-pool3-capture2	0.6826	0.0017	412	780	343	537	777	1017	4986
Line B-pool4-capture2	0.7065	0.0017	427	1000	474	650	984	1332	6934
Line C-pool1-capture2	0.5488	0.0017	331	742	329	515	745	967	18073
Line C-pool2-capture2	0.5997	0.0017	362	855	379	588	859	1121	10325
Line C-pool3-capture2	0.6308	0.0017	381	807	353	564	808	1043	13608
Line C-pool4-capture2	0.6183	0.0017	373	831	377	557	828	1097	7709
Line D-pool1-capture2	0.6135	0.0017	370	833	357	579	828	1080	4380
Line D-pool2-capture2	0.6414	0.0017	387	991	420	690	988	1291	6173
Line D-pool3-capture2	0.6530	0.0017	394	897	386	622	891	1164	5379
Line D-pool4-capture2	0.6482	0.0017	391	1325	590	912	1304	1705	19930
Line D-pool5-capture2	0.6360	0.0017	384	864	375	598	856	1117	5537
Line D-pool6-capture2	0.6648	0.0017	402	900	403	610	890	1176	13525

Table 3.23: Summary of capture statistic for capture set 2 and the different pools for each line. The statistics reported are fr: fraction of reads covering the target and ft fraction of reads covering the genome. Enrichiment is the ratio: fr/ft. Coverage metrics are: *average coverage:* average coverage for the target bases and *sd coverage:* coverage standard deviation. The last 4 columns display the 25th, 50th (median) , 75th and 100th percentile for coverage.

Sample Name	fr	ft	enrichment	average coverage	sd_coverage	25%	50%	75%	100%
Line H-pool1-capture3	0.6938	0.0019	358	839	486	596	828	1038	12441
Line H-pool2-capture3	0.7297	0.0019	377	636	377	446	626	788	9688
Line H-pool3-capture3	0.7130	0.0019	368	753	438	531	739	934	12048
Line H-pool4-capture3	0.7083	0.0019	366	566	345	405	555	692	9254
Line H-pool5-capture3	0.7004	0.0019	362	766	435	543	749	945	12445
Line H-pool6-capture3	0.7328	0.0019	379	714	380	501	706	893	9312
Line B-pool1-capture3	0.7142	0.0019	369	651	250	474	661	828	3351
Line B-pool2-capture3	0.6974	0.0019	360	1019	381	755	1034	1286	4912
Line B-pool3-capture3	0.7187	0.0019	371	423	159	312	428	535	2135
Line B-pool4-capture3	0.7232	0.0019	374	679	262	491	690	867	3392
Line C-pool1-capture3	0.7161	0.0019	370	1084	433	795	1093	1362	7288
Line C-pool2-capture3	0.7007	0.0019	362	700	285	502	705	889	3987
Line C-pool3-capture3	0.6615	0.0019	342	870	347	629	873	1103	4277
Line C-pool4-capture3	0.7210	0.0019	373	768	306	553	776	976	3978
Line D-pool1-capture3	0.7083	0.0019	366	1018	457	740	1014	1269	9618
Line D-pool2-capture3	0.7088	0.0019	366	751	360	548	750	932	8439
Line D-pool3-capture3	0.7082	0.0019	366	637	293	463	637	794	6568
Line D-pool4-capture3	0.6961	0.0019	360	669	377	480	656	822	9851
Line D-pool5-capture3	0.6907	0.0019	357	797	563	575	780	963	15583
Line D-pool6-capture3	0.6979	0.0019	361	628	375	454	620	769	10076

Table 3.24: Summary of capture statistic for capture set 3 and the different pools for each line. The statistics reported are fr: fraction of reads covering the target and ft fraction of reads covering the genome. Enrichiment is the ratio: fr/ft. Coverage metrics are: *average coverage:* average coverage for the target bases and *sd coverage:* coverage standard deviation. The last 4 columns display the 25th, 50th (median) , 75th and 100th percentile for coverage.

Chromosome coverage bar plot for the reads mapped and the targeted regions are represented in figures 3.29, 3.30, 3.31 for one sample (line B pool1) for the 3 capture sets. The same plot for the other samples are provided in supplementary data. The plots show the fraction of total reads mapping to each chromosome and the fraction of reads mapping to the targetted region for each chromosome. The fraction of reads on target and reads mapped to chromosome are not always the same but are closely correlated. Some chromosomes have reads mapping to them while there are no targetted reads on them, suggesting some off target effects. There are also reads mapping to regions that were not on the capture list. This is expected given the fraction of reads on target shown in the summary tables, depending on the capture set a large fraction of reads might be off target, see section 4.2.4 for the discussion about capture sets.

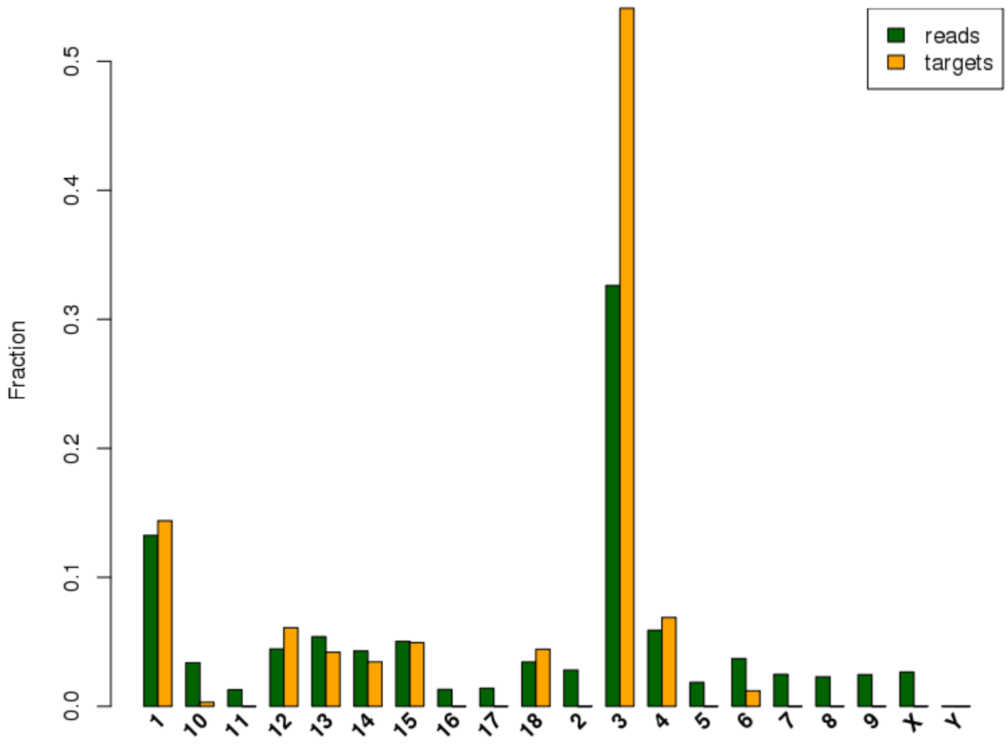


Figure 3.26: Capture 1 chromosome bar plot for line B pool . The green bars represent the reads fraction mapped to each chromosome and in yellow are the read fraction mapped to targetted regions.

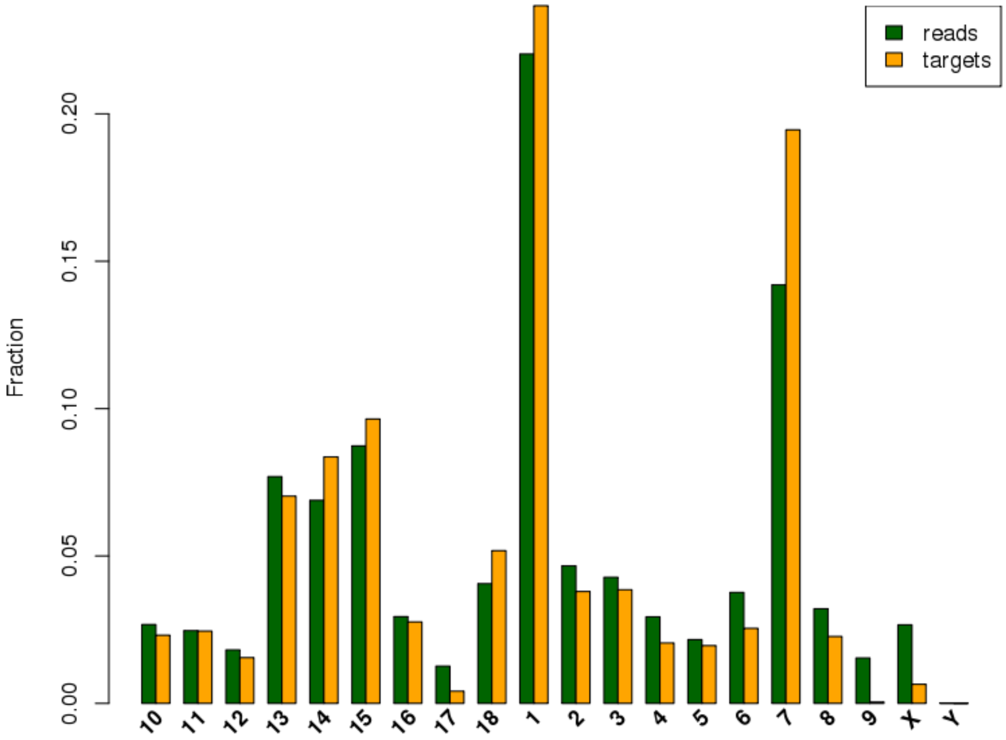


Figure 3.27: Capture 2 chromosome bar plot for line B pool1. The green bars represent the fraction of reads mapped and in yellow are the fraction of targetted regions.

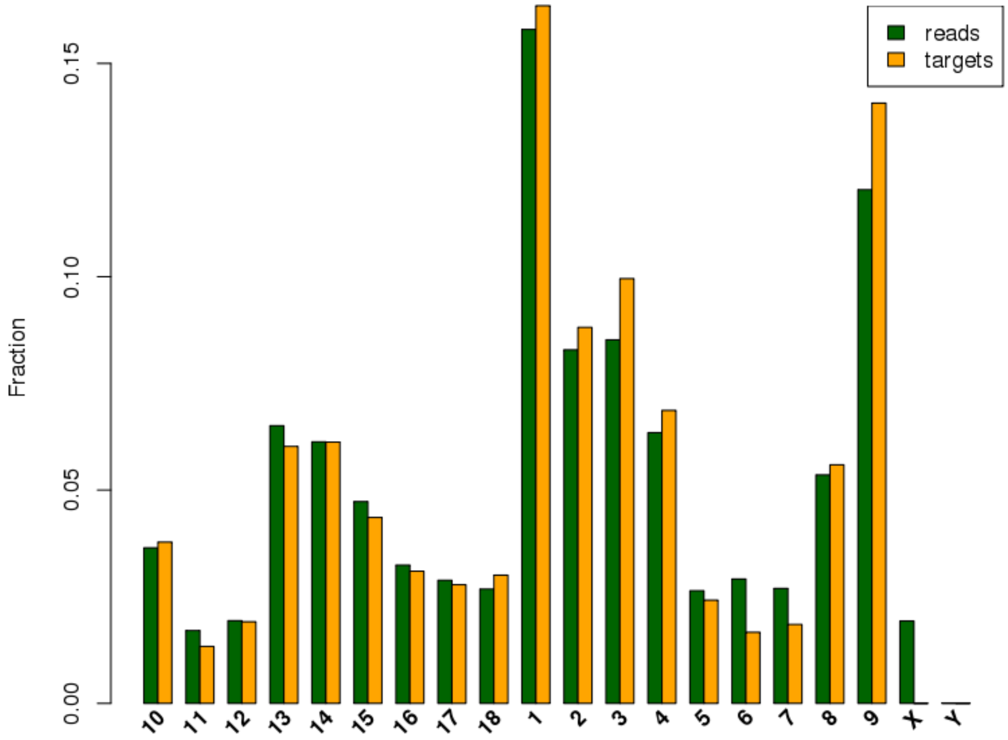


Figure 3.28: Capture 3 chromosome bar plot for line B pool1, The green bars represent the reads mapped and in yellow are the targetted regions.

Coverage distributions for reads on target for pool 1, line B are shown in figures 3.29,3.30,3.31 for capture set 1, 2 and 3. For the other pools and lines the plots are in supplemental material. The threshold for good coverage was set at 20x (20 bases depth for a given location). All of the capture sets achieve this for a large fraction of the target base coverage. Coverage was generally very high as seen in tables 3.22, 3.23 and 3.24. Capture set 1 had the highest coverage of all of the capture sets and capture set 3 the lowest. According to figure 3.29, capture 1 had more than 50% of target based covered at more than 1500x, only a very small fraction of the regions, around one percent, are covered at less than 20X. The highest coverage for a fraction of the regions reached more than 3500X. For capture set 2, 50% or more of the target regions were covered by at least 500X, again only a small fraction of the regions, around one percent, was covered by less than 20X. The highest coverage was above 2000X. Finally for capture 3 the coverage was not as deep but remained very high. Half of the targetted regions are covered by more than 600X, while the upper end of the coverage is lower compared to the other two capture sets, it reached 1300x. At the lower end of the coverage, a slightly larger proportion of the target regions were covered with less than 20X, but it remains a very small proportion, around two to three percent.

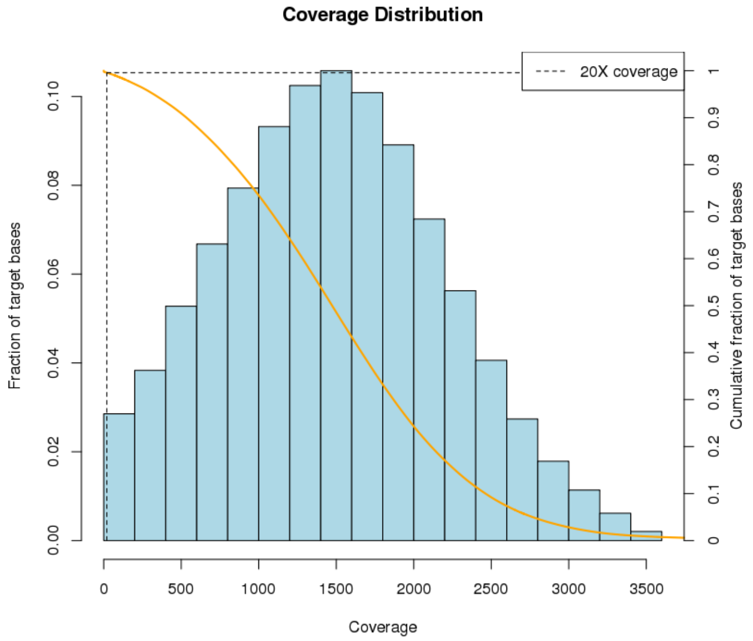


Figure 3.29: Coverage distribution for capture 1 line B pool1. The blue histogram shows the distribution of coverage, the y-axis for this being on the left side. The yellow line shows the coverage relative to the cumulative fraction of target bases, the y axis being on the right side. The dotted line shows the target 20x coverage.

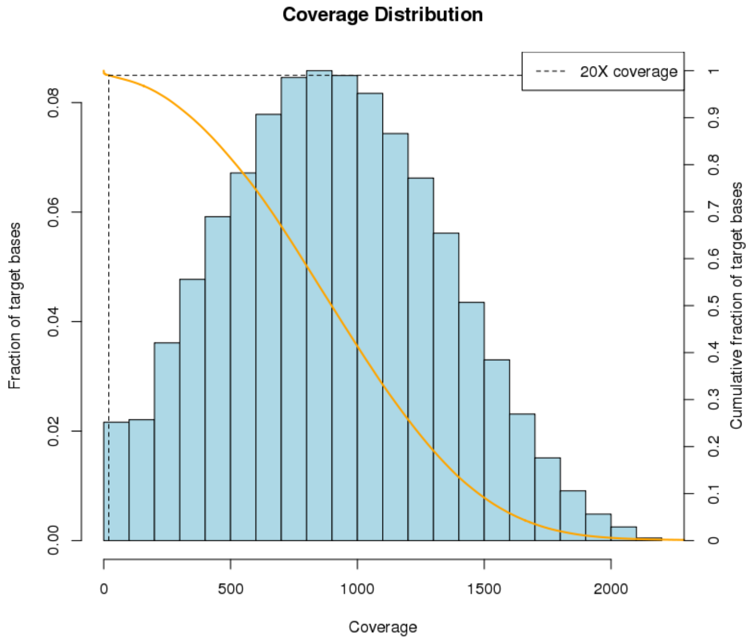


Figure 3.30: Coverage distribution for capture 2 line B pool1. The blue histogram shows the distribution of coverage, the y-axis for this being on the left side. The yellow line shows the coverage relative to the cumulative fraction of target bases, the y axis being on the right side. The dotted line shows the target 20x coverage.

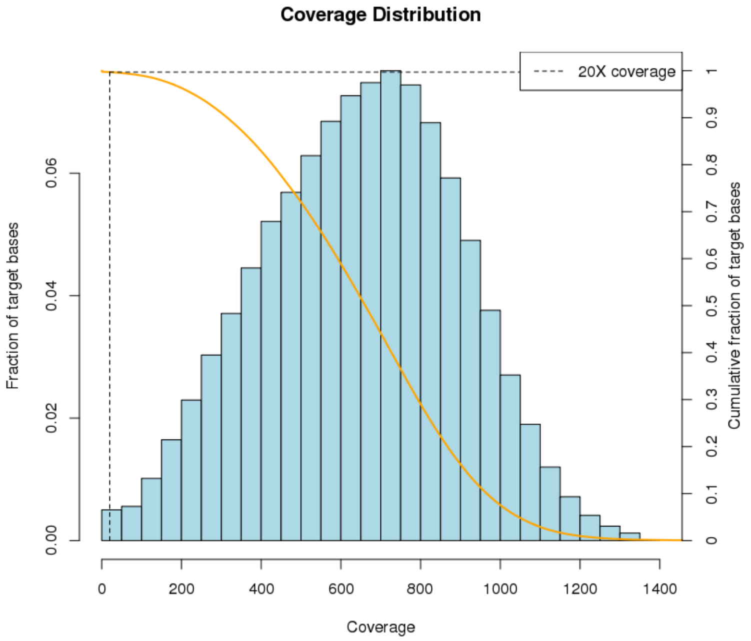


Figure 3.31: Coverage distribution for capture 3 line B pool 1. The blue histogram shows the distribution of coverage, the y-axis for this being on the left side. The yellow line shows the coverage relative to the cumulative fraction of target bases, the y axis being on the right side. The dotted line shows the target 20x coverage.

Normalised coverage plots for pool 1 line B are shown for all capture sets on figures 3.32, 3.33 and 3.34. The normalised coverage is calculated by taking the per base coverage and dividing it by the average coverage for all targetted bases. As it is not dependent on the number of reads, it makes it easier to compare data between samples. The graphs show that the coverage is uniform and behaves similarly for all the capture sets. More than 80 percent of the data is covered by at least half of the average coverage value for all of the capture sets.

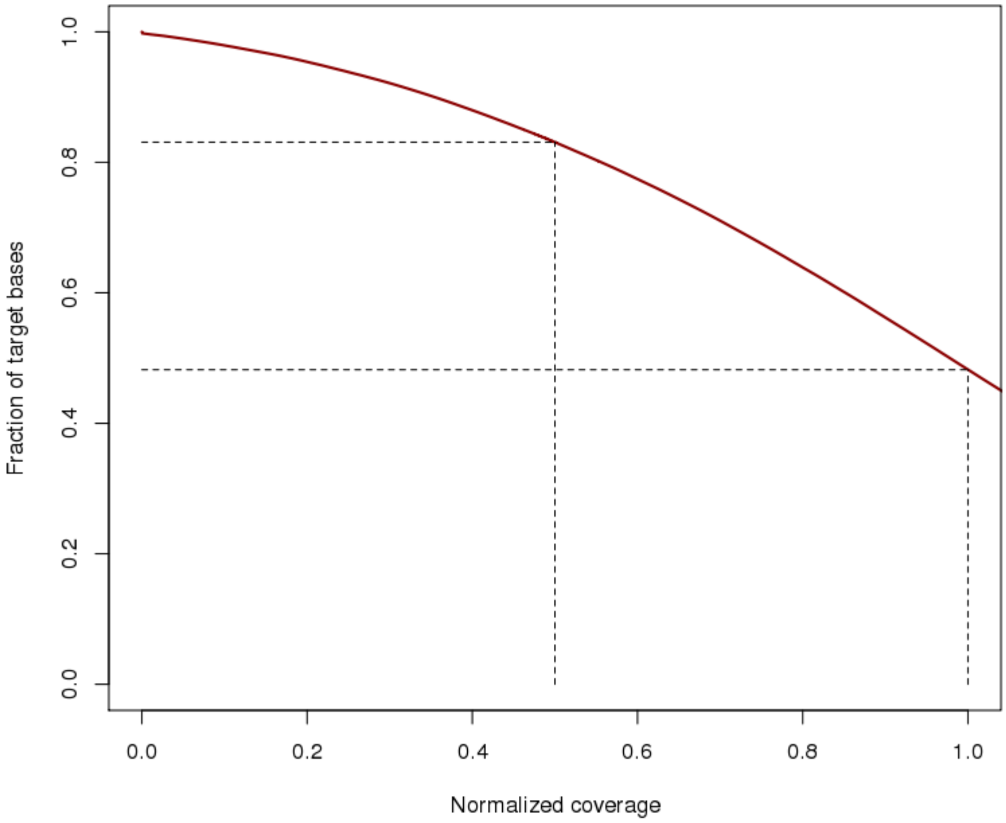


Figure 3.32: Normalised coverage plot for capture set 1, pool 1 line B.The y axis represents the fraction of target base covered and the x axis the normalised coverage. The normalised coverage is calculated by dividing the coverage of a given base by the average coverage of the whole sample. At 0.5 the fraction of bases covered will be covered by at least half of the average while at 1 it will be covered by at least the average.

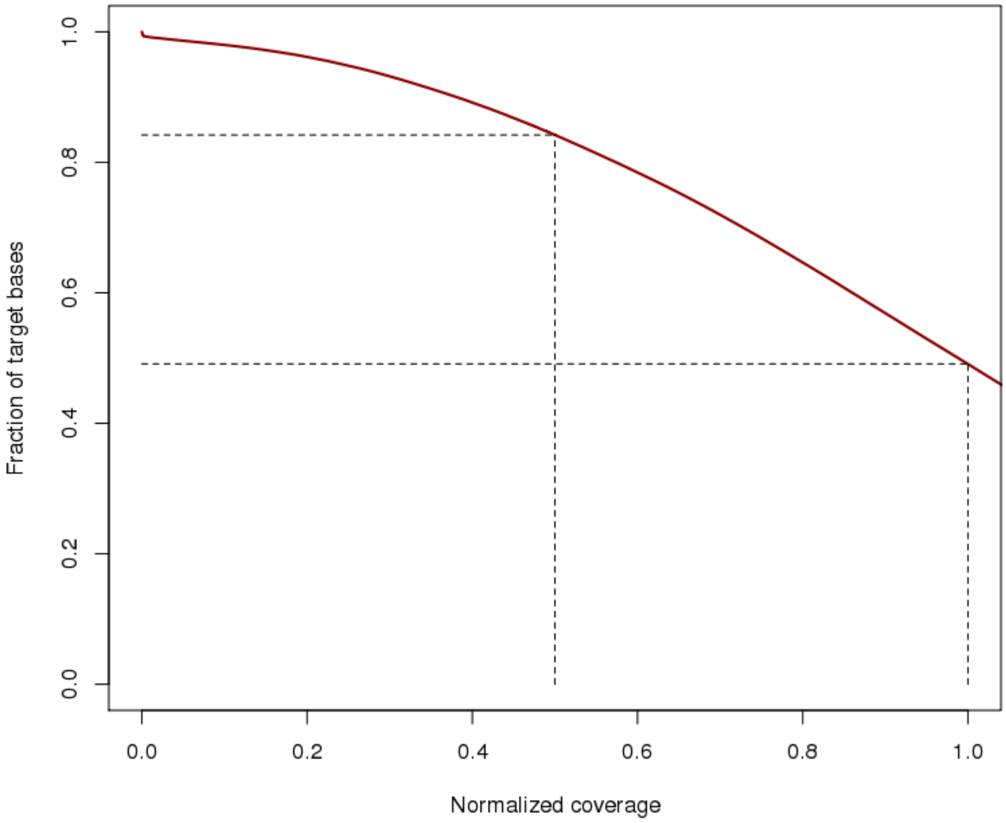


Figure 3.33: Normalised coverage plot for capture set 2, pool1 line B. The y axis represents the fraction of target base covered and the x axis the normalised coverage. The normalised coverage is calculated by dividing the coverage of a given base by the average coverage of the whole sample. At 0.5 the fraction of bases covered will be covered by at least half of the average while at 1 it will be covered by at least the average.

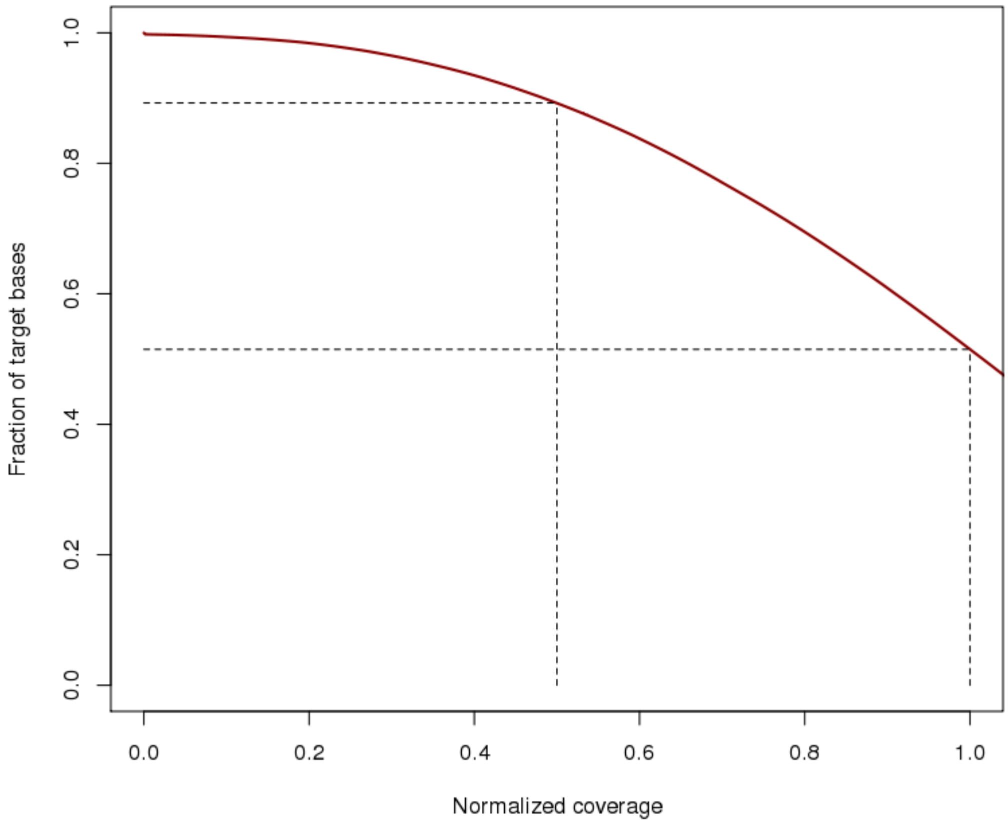


Figure 3.34: Normalised coverage plot for capture set 3, pool1 line B. The y axis represents the fraction of target base covered and the x axis the normalised coverage. The normalised coverage is calculated by dividing the coverage of a given base by the average coverage of the whole sample. At 0.5 the fraction of bases covered will be covered by at least half of the average while at 1 it will be covered by at least the average.

3.3 Mapped reads processing and variants analysis.

The GATK part of the analysis pipeline produced variants files for each sequencing pool that was processed. The GATK pipeline is described in details in section 2.10.

3.3.1 Quality score recalibration

The graph below shows the results of applying the quality score recalibration on pool1 of line B for capture 1: the other files for the other capture sets are available in supplementary data. The accuracy plot shows the difference between the empirical value and the reported quality score, negative values meaning that the the reported score is higher than the empirical one. Figure 3.35 shows that the quality score before correction is higher than the one after correction.. In figure 3.36 almost all of the graphs show that the quality score before recalibration was higher than the one after recalibration. The spread of the scores was also smaller before adjustment. Only when looking at the mean quality score can we see that in some contexts the quality score before recalibration was lower.

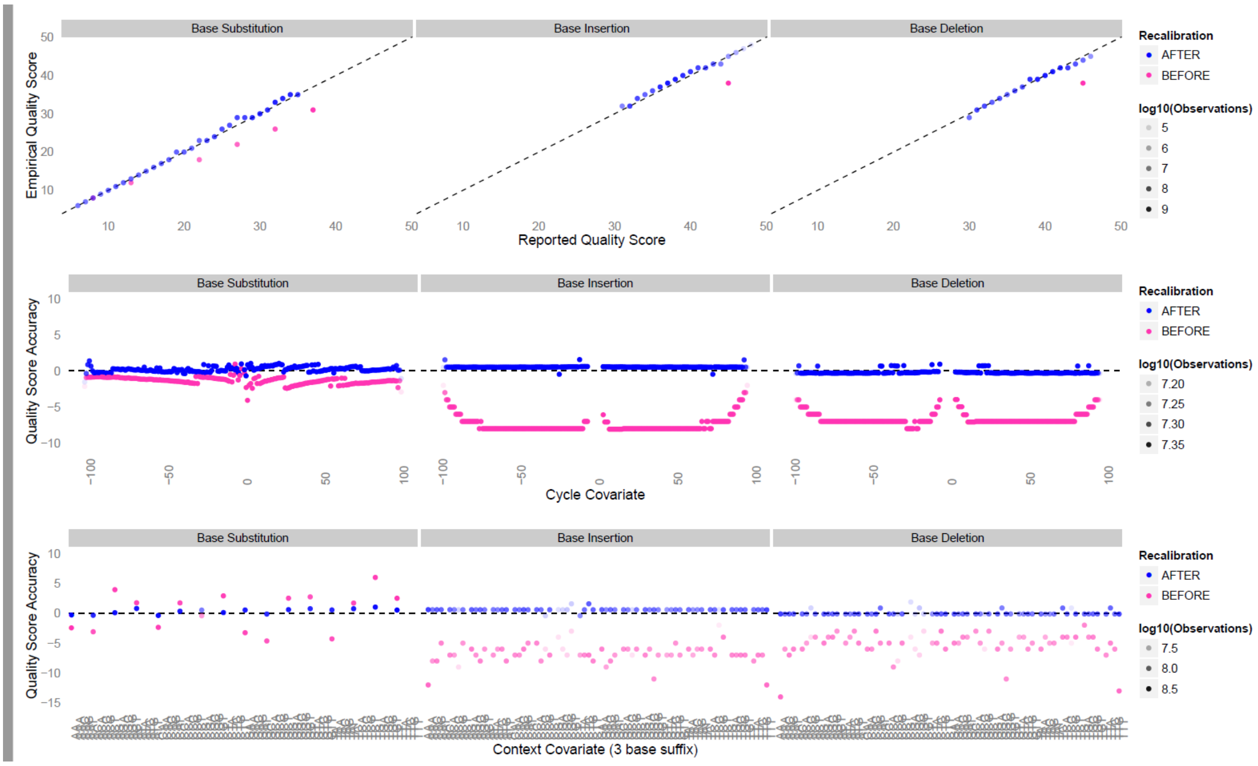


Figure 3.35: Quality score before and after recalibration. Each column of the graph represent a potential deviation from the reference sequence. In order: base substitution, base insertion and base deletion. The first row displays the graph for the empirical score (calculated using machine learning) against reported quality score. The second row displays the quality score accuracy according to the cycle of the reads, the positive values on the X axis being the first read and the negative values being the second read. The Y axis represents the accuracy score calculated by subtracting the reported (observed) quality score from the empirical score, a negative value representing over confident quality score. The last row display the quality score accuracy according to the context covariate which are the 3 bases before it. All of the of the data suggest that the reported quality score is higher than it should be.

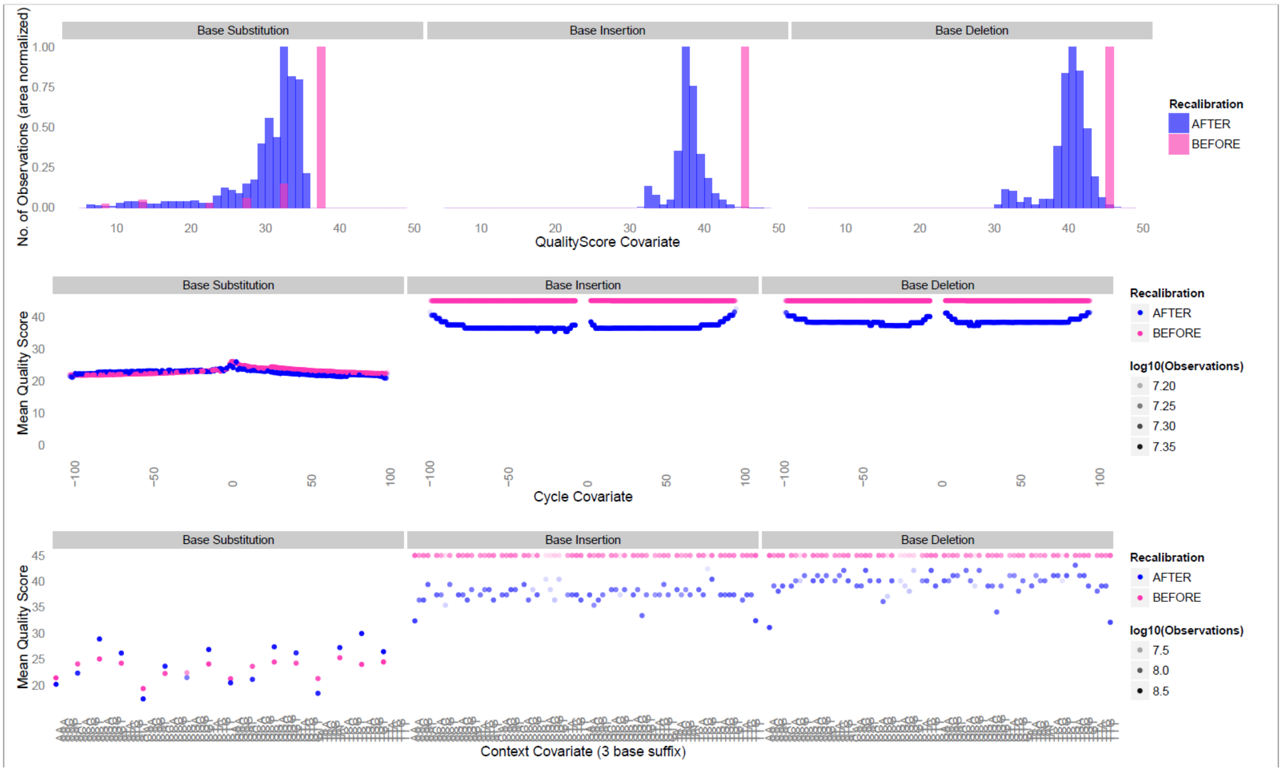


Figure 3.36: Quality scores before and after recalibration. Each column relates to a particular event, in order: base substitution, base insertion and base deletion. The first row shows the distribution of the quality score before and after recalibration. The second row, the mean quality score before and after in relation to the read cycle, positive values on the X axis representing the first read, negative values the second read. The third row shows the mean quality score in relation to the 3 bases preceding the base of interest.

3.3.2 Variant calling

The number of variants called for each capture set and each pool before and after filtering is given in table 3.25. The numbers in this table were computed after the variant discovery was completed.

Table 3.27 contains the number of variants after grouping per category as described in the material and methods, see section 2.9.1. The numbers are calculated from the updated table with variant remapped to build 11 of the genome, see section 2.11.3 for more details about updating the coordinates to build 11. A unique variant corresponds to a variant with a unique position in the genome: variants sharing the same location are removed from the counting.

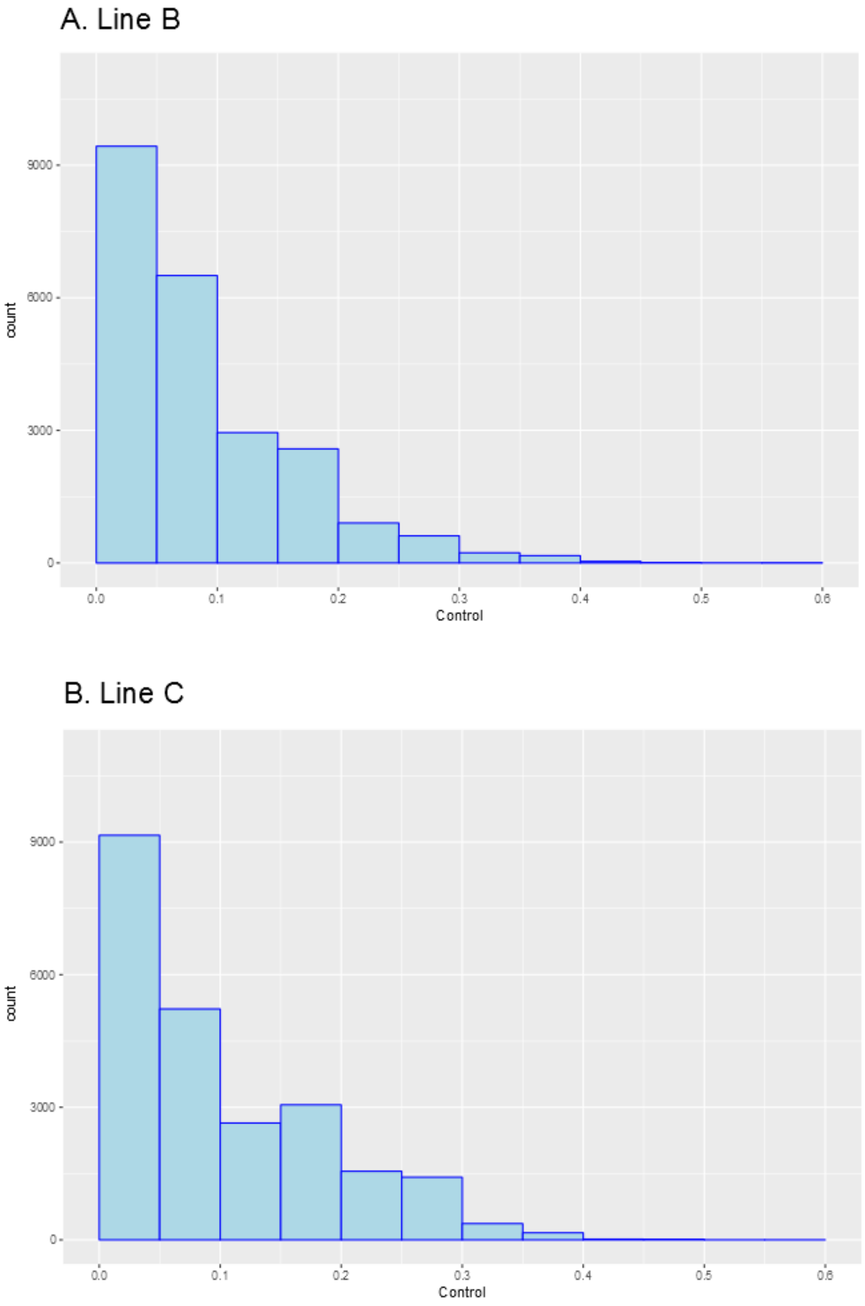
Figures 3.37, show the density distribution of the difference in allele frequencies between the control pools and between the control pools and infanticide pools. The figure (a) displays the difference in allele frequencies between the control for each capture set, grouping all the variants from all the lines. The figure (b) plots the difference in allele frequencies for each of the capture sets and each category within that set. For both figures the vertical lines show mean plus 3 standard deviations, for each capture set for the first figure and for each capture set and category for the second one.

Figures 3.38, 3.39, 3.40, 3.41, 3.42 and 3.43 display the histogram of the repartition of the difference in allele frequency between the control pool for each capture set and each line.

Table 3.26 shows the results of a mock selection of the variant based on the control pools.

Tables 3.28, 3.29, 3.30, 3.31, 3.32, 3.33, 3.34, 3.35, 3.36 show the percentage of selected variants using the filtering criteria to identify variants with difference in allele frequencies between control and infanticide pools. The percentage is calculated after the variant coordinates have been updated to the version 11 of the genome. The numbers are similar for the same category between the capture sets.

All the variant files used to generate these graphs and tables are available in the supplementary material attached to this thesis.



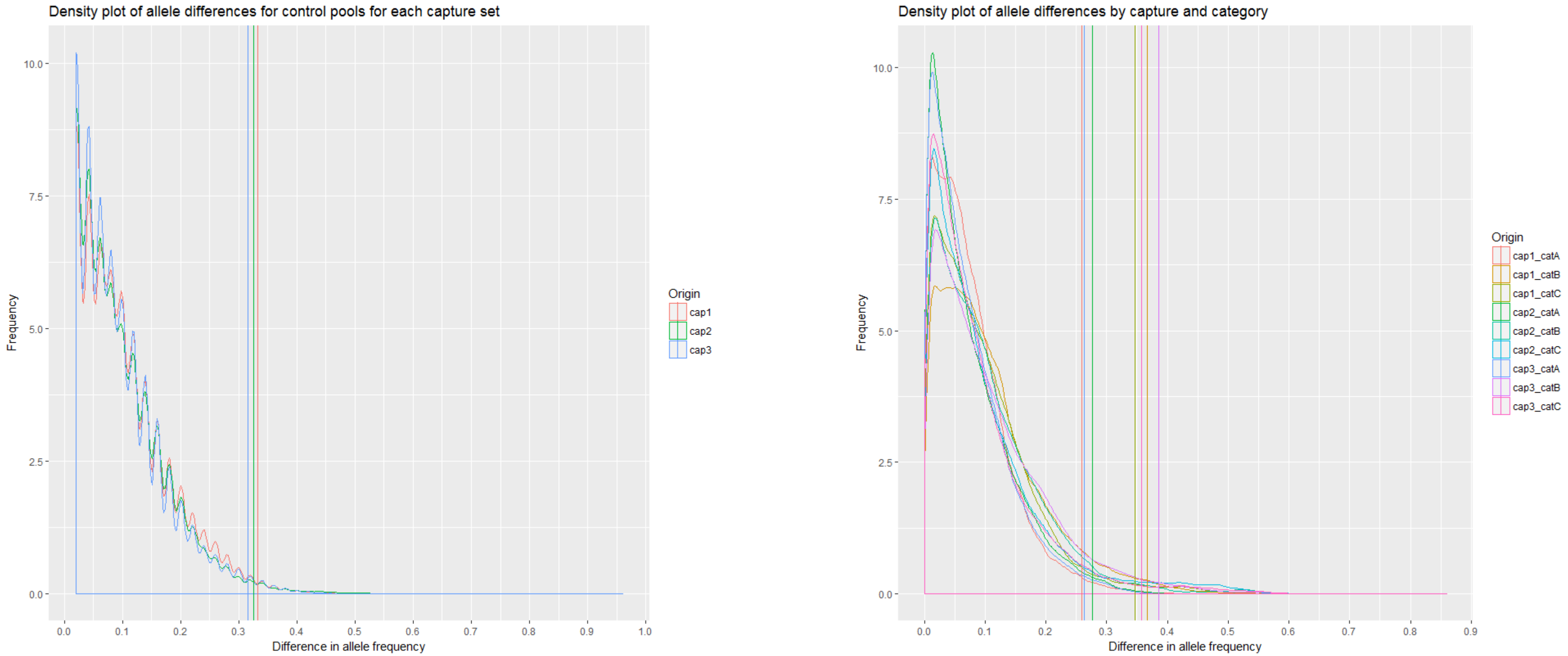
A. Line B

B. Line C

Figure 3.38: Histogram of within control pools allele frequency difference for line B and C capture set one. X axis displays the difference in allele frequency between the control pools. Y axis: “count” refer to the number of SNPs.

Line and Pool	Raw variants set 1	Filtered variants set 1	Raw variants set 2	Filtered variants set 2	Raw variants set 3	Filtered variants set 3
Line B pool 1	27576	20771	23856	19768	23726	20125
Line B pool 2	29256	19964	27466	19823	43695	34371
Line B pool 3	39768	20825	33270	19689	53441	36880
Line B pool 4	39226	21108	32338	20442	51557	35809
Line C pool 1	35687	20529	31249	19520	55366	37096
Line C pool 2	26955	19388	26563	19809	43066	33499
Line C pool 3	39712	20772	33426	20354	55602	36323
Line C pool 4	40942	20547	33927	19235	57111	36895
Line D pool 1	40135	23302	32555	20068	50442	30219
Line D pool 2	38708	25771	31902	20187	47046	32149
Line D pool 3	41403	25556	32410	19139	49588	33530
Line D pool 4	40958	23159	32320	20181	48079	31311
Line D pool 5	41759	23561	31533	19178	49629	32368
Line D pool 6	41031	24948	32407	19525	51289	32200
Line H pool 1	31590	18736	28156	18020	47757	33257
Line H pool 2	35363	21013	28047	17159	48418	31620
Line H pool 3	31613	16408	28173	18251	46925	30785
Line H pool 4	33264	21389	26968	18569	45146	31852
Line H pool 5	37277	22054	31929	17732	54711	32268
Line H pool 6	37594	22724	31038	18367	54386	32753

Table 3.25: Variants called for Capture set 1,2 and 3. The table shows the raw and filtered variant number for each capture set.



(a) Density plot of the average difference between the allele of the control pool

(b) Density plot of the average difference between the infanticide pools against the control pools for all the variants compared

Figure 3.37: Sub Figure a: Density plot of the average difference between the allele of the control pools, for each capture set for all of the variants. The vertical lines show the threshold of mean plus three stand deviation for each of the capture set. Cap 1: capture set 1, cap 2: capture set 2, cap 3: capture set 3.
Sub Figure b: Density plot of the average difference between the infanticide pools against the control pools for all the variants compared. Vertical lines show the threshold of mean plus three standard deviation for each of the category and each of the capture set. Cap1_catA: capture set 1, category A. Cap1_catB: capture set 1, category B. Cap1_catC: capture set 1, category C. Cap2_catA: capture set 2, category A. Cap2_catB: capture set 2, category B. Cap2_catC: capture set 2, category C. Cap3_catA: capture set 3, category A. Cap3_catB: capture set 3, category B. Cap3_catC: capture set 3, category C.

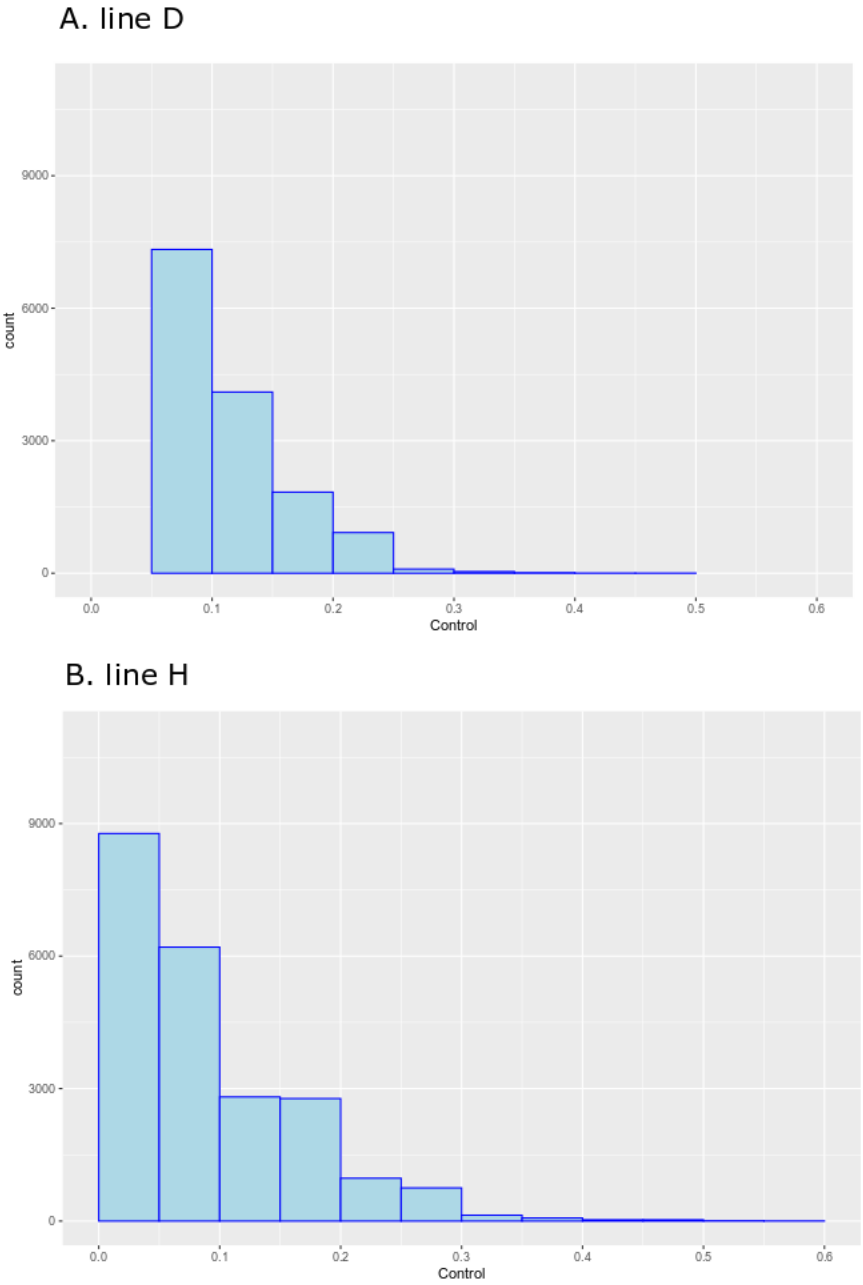


Figure 3.39: Histogram of within control pools allele frequency difference for line D and H capture set one. X axis displays the difference in allele frequency between the control pools. Y axis: "count" refer to the number of SNPs.

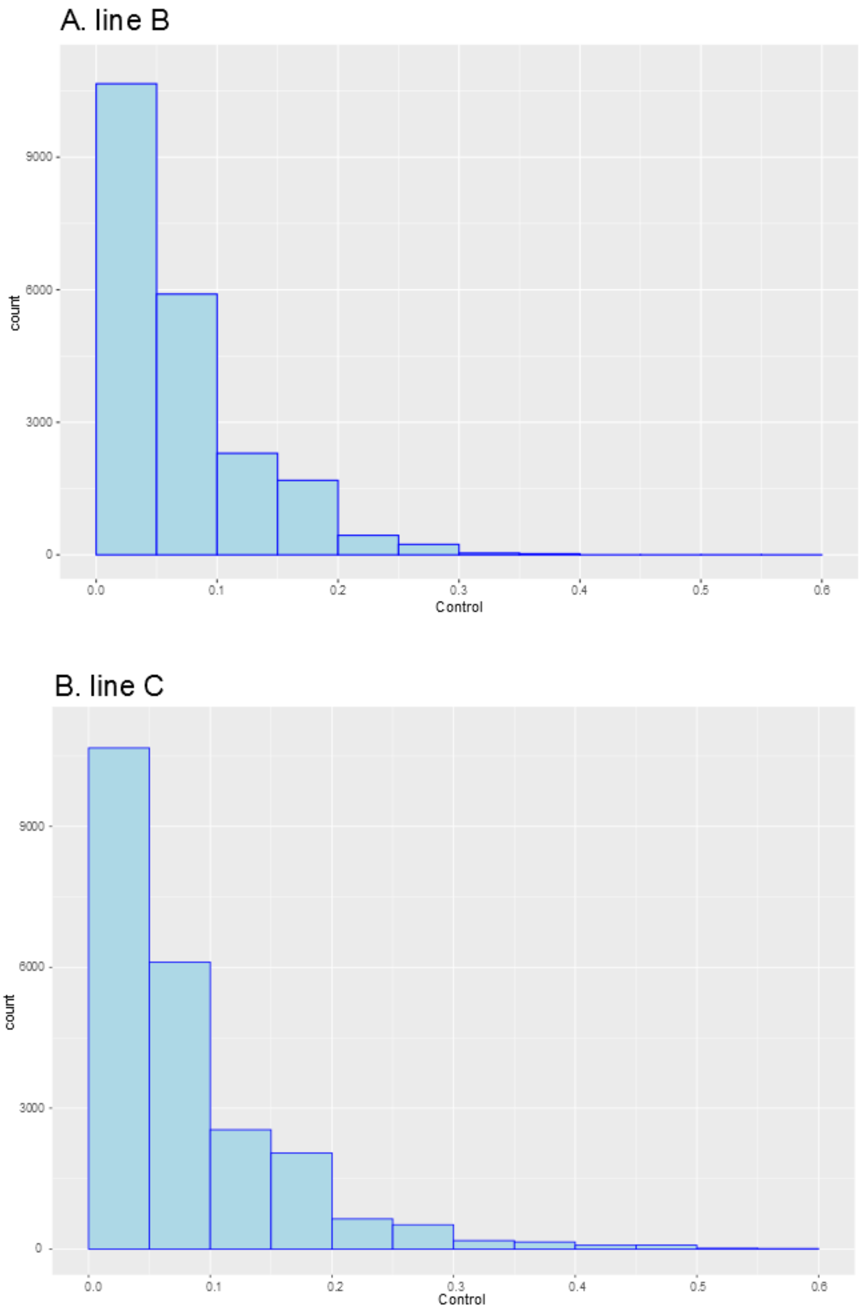


Figure 3.40: Histogram of within control pools allele frequency difference for line B and C capture set 2. X axis displays the difference in allele frequency between the control pools. Y axis: “count” refer to the number of SNPs.

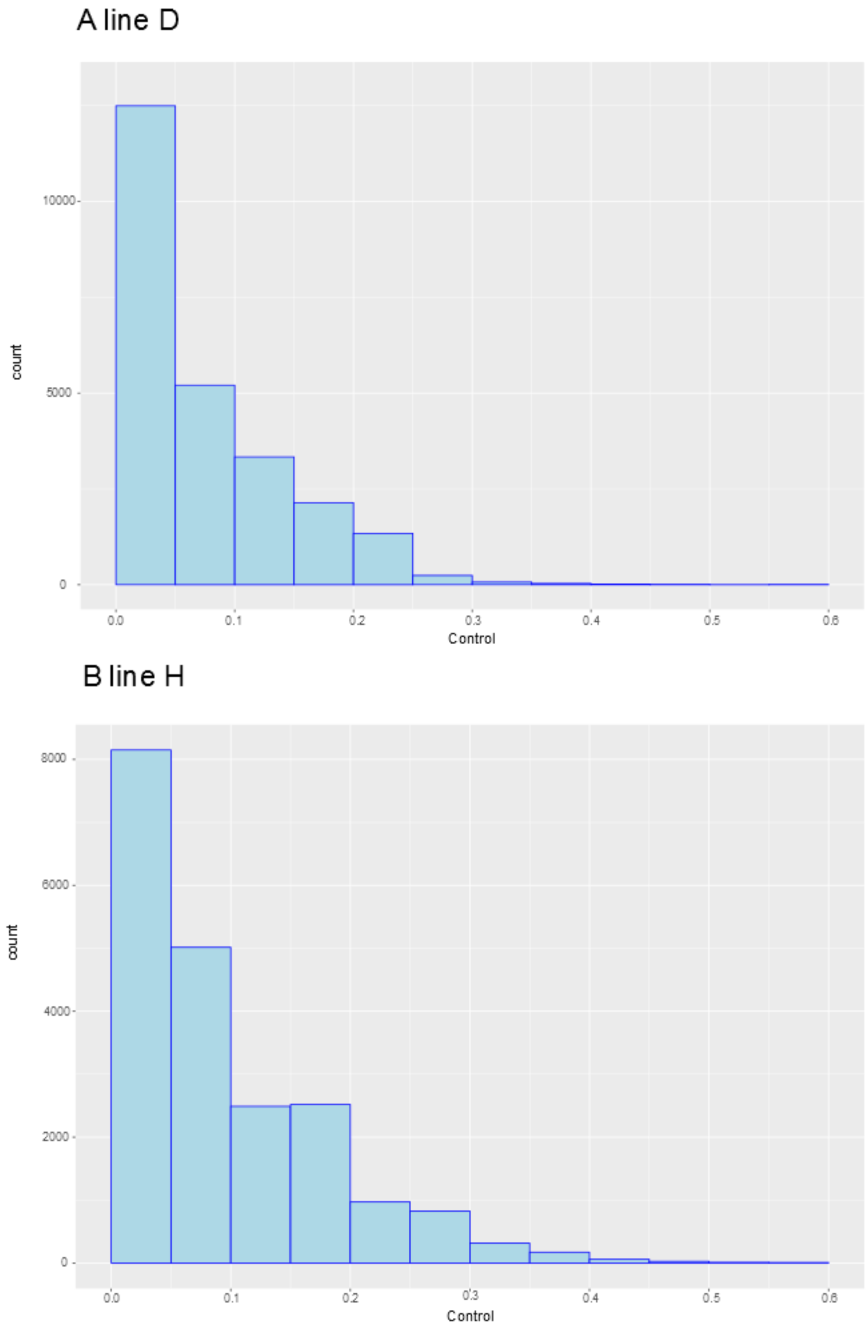


Figure 3.41: Histogram of within control pools allele frequency difference for line D and H capture set 2. X axis displays the difference in allele frequency between the control pools. Y axis: "count" refers to the number of SNPs.

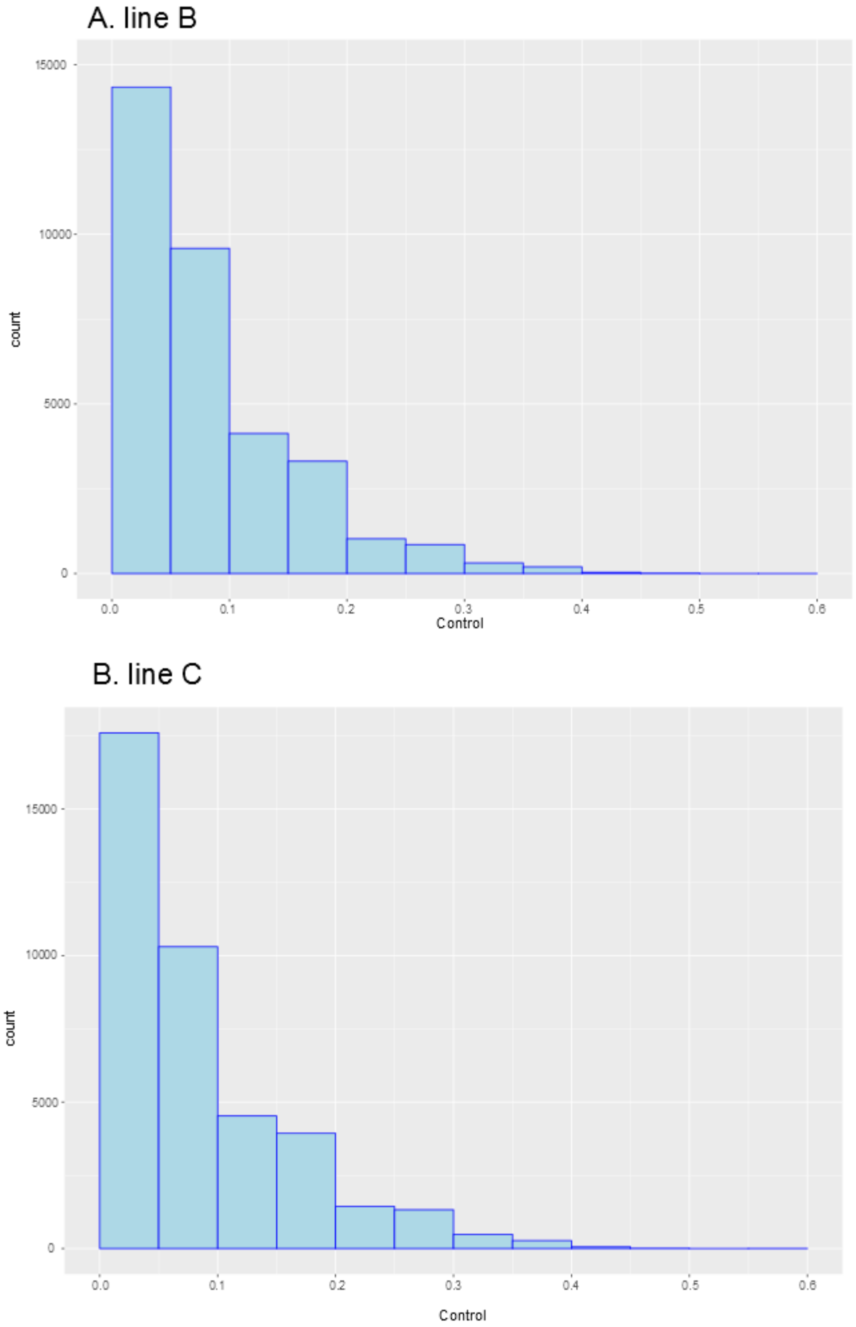


Figure 3.42: Histogram of within control pools allele frequency difference for line B and C capture set 3. X axis displays the difference in allele frequency between the control pools. Y axis: “count” refer to the number of SNPs.

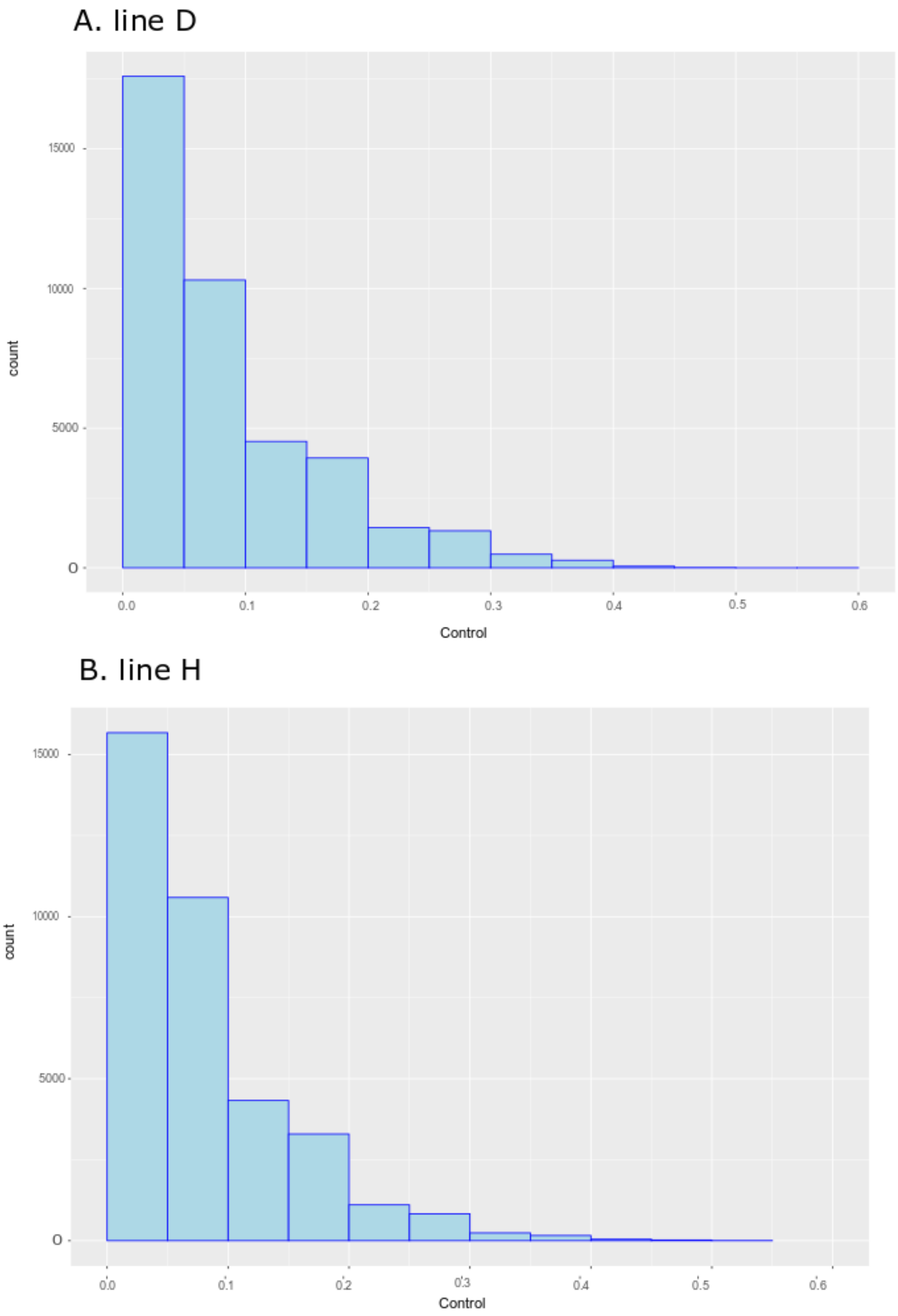


Figure 3.43: Histogram of within control pools allele frequency difference for line D and H capture set 3. X axis displays the difference in allele frequency between the control pools. Y axis: "count" refer to the number of SNPs.

Capture set	Total unique variants	Selected (mean + 3sd)	Percentage selected from total
Capture 1	30660	867	2.8%
Capture 2	26231	828	3.1%
Capture 3	48441	1932	4%

Table 3.26: Results of the mock selection of the variants on the control pools for each capture set. The total number of variant can include variant duplicates found in several lines. Selection is done using the mean of the allele difference for all the line in the capture set and adding 3 standard deviation of the allele difference.

Pools	Unique variants	Threshold Mean + 3 sd	Number of target SNPs	Percentage of total
Capture set 1 category A	27714	0.2617	1067	3.8
Capture set 1 category B	28738	0.3692	1041	3.6
Capture set 1 category C	28046	0.3455	1175	5
Capture set 2 category A	22872	0.2773	782	3.4
Capture set 2 category B	23343	0.3564	966	4.1
Capture set 2 category C	23172	0.3760	1511	6.5
Capture set 3 category A	42855	0.2645	1307	3
Capture set 3 category B	44175	0.3878	1894	4.3
Capture set 3 category C	43327	0.3576	2068	4.7

Table 3.27: Variants after grouping into categories and target identified for each category.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	1049	0	0
3	13771	160	1.16
4	2828	171	6.05
6	416	19	4.57
10	149	5	3.36
12	2482	198	7.98
13	894	4	0.45
14	890	246	27.64
15	1648	105	6.37
18	1617	58	3.59

Table 3.28: Target capture overlap summary, Capture set 1 category A. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variant in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	1087	1	0.09
3	15496	423	2.73
4	2981	284	9.53
6	467	2	0.43
10	149	0	0
12	2911	96	3.3
13	958	3	0.31
14	899	69	7.68
15	1710	49	2.87
18	1630	15	0.92

Table 3.29: Target capture overlap summary, Capture set 1 category B. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variant on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	1062	0	0
3	14239	334	2.35
4	2939	49	1.67
6	431	236	54.76
10	148	1	0.68
12	2824	20	0.71
13	952	311	32.67
14	920	153	16.63
15	1679	49	2.92
18	1631	8	0.49

Table 3.30: Target capture overlap summary, Capture set 1 category C. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	5203	11	0.21
2	1079	4	0.37
3	769	0	0
4	1645	0	0
5	1244	6	0.48
6	1050	0	0
7	3553	69	1.94
8	465	57	12.26
9	1554	1	0.06
10	912	18	1.97
11	1121	28	2.5
12	792	16	2.02
13	3364	44	1.31
14	3128	9	0.29
15	1661	2	0.12
16	1141	74	6.49
17	42	0	0
18	1049	6	0.57

Table 3.31: Target capture overlap summary, Capture set 2 category A. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	3706	280	7.56
2	1164	6	0.52
3	821	113	13.76
4	860	1	0.12
5	1349	63	4.67
6	888	7	0.79
7	3975	15	0.38
8	474	4	0.84
9	13	0	0
10	969	69	7.12
11	1505	6	0.4
12	638	50	7.84
13	2483	170	6.85
14	2060	65	3.16
15	676	40	5.92
16	502	3	0.6
17	54	1	1.85
18	953	1	0.1

Table 3.32: Target capture overlap summary, Capture set 2 category B. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	3679	126	3.42
2	1160	0	0
3	818	2	0.24
4	863	0	0
5	1391	3	0.22
6	884	139	15.72
7	3870	80	2.07
8	475	22	4.63
9	12	1	8.33
10	964	51	5.29
11	1529	645	42.18
12	658	43	6.53
13	2405	300	12.47
14	1962	4	0.2
15	631	23	3.65
16	490	11	2.24
17	48	0	0
18	925	16	1.73

Table 3.33: Target capture overlap summary, Capture set 2 category C. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	4285	63	1.47
2	3851	221	5.74
3	3538	46	1.3
4	2809	156	5.55
5	1304	2	0.15
6	987	18	1.82
7	202	0	0
8	2783	121	4.35
9	5298	68	1.28
10	1763	168	9.53
11	804	59	7.34
12	908	27	2.97
13	2751	40	1.45
14	2220	5	0.23
15	2081	8	0.38
16	1327	98	7.39
17	911	53	5.82
18	1991	72	3.62

Table 3.34: Target capture overlap summary, Capture set 3 category A. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	4495	212	4.72
2	4003	165	4.12
3	3906	32	0.82
4	3057	488	15.96
5	1403	34	2.42
6	1041	1	0.1
7	219	0	0
8	2847	51	1.79
9	6325	222	3.51
10	1833	151	8.24
11	832	4	0.48
12	1063	2	0.19
13	2873	190	6.61
14	2415	9	0.37
15	2322	10	0.43
16	1419	56	3.95
17	1123	34	3.03
18	2122	157	7.4

Table 3.35: Target capture overlap summary, Capture set 3 category B. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

Target chromosome	Overlap capture set and variants	Overlap capture set selected variant	Percentage overlap target to all variants
1	4319	41	0.95
2	4036	356	8.82
3	3697	209	5.65
4	2807	1	0.04
5	1193	0	0
6	988	4	0.4
7	200	0	0
8	2584	0	0
9	4944	2	0.04
10	1801	734	40.76
11	825	5	0.61
12	957	58	6.06
13	2857	157	5.5
14	2346	198	8.44
15	2153	13	0.6
16	1416	29	2.05
17	954	20	2.1
18	2128	129	6.06

Table 3.36: Target capture overlap summary, Capture set 3 category C. For each chromosome the table shows: the number of variants identified within the capture intervals, the number of selected variants in the capture intervals, and lastly the percentage of selected variants to all variants on the chromosome.

3.3.3 SNP of interest

Tables 3.37 and 3.38 display the SNPs identified using the variable threshold that were identified as missense according to the SIFT prediction. The large majority of this SNP are found on line B and C, the two dam lines. The number of SNP identified in each capture set goes from 8 for capture set 1 to 12 and 15 for capture set 2 and 3 respectively. Two SNPs (rs332900783 and rs 81209170) were found in common between two capture sets, capture set 2 and three for category A. The highest allele frequency different between the pools for these SNPs is 58.7% and the lowest 26.6%. To assess the potential impact of the SNPs, they were checked using a manual approach (see section 2.11.4).

3.3.4 Capture set 1 examples of regions of interest.

The region of interest were selected for capture set 1 as region with a large number of SNPs present and covering gene or region of interest. Figures 3.44 and 3.45 show two examples of regions of interest selected for capture set 1 category B and C. The rest of the selected region are discussed in section 4.3.6 and the genomic plots provided in supplementary material.

ID	CHR	POS	Average.diff	Allele.Frequency	REF	ALT	sift_prediction	sift_score	Origin	Line	Gene
rs337246460	1	81891724	0.275(A), 0.4(B)	Agg : 0.250(B),0.500(C), Norm : 0.720,0.580,	A	G	NA	NA	Cap3 cat A, B	C	*TSPYL4*
rs321068974	1	81892394	0.43	Agg : 0.250, Norm : 0.720,0.640,	T	C	tolerated	1	Cap3 cat B	C	*TSPYL4*
rs80970369	1	81892966	0.41	Agg : 0.250, Norm : 0.720,0.600,	G	T	tolerated - low confidence	0.32	Cap3 cat B	C	*TSPYL4*
rs327767403	1	81908800	0.39	Agg : 0.250, Norm : 0.700,0.580,	A	G	tolerated	0.23	Cap3 cat B	C	*TSPYL1*
rs325683130	4	109665261	0.425	Agg : 0.875, Norm : 0.520,0.380,	C	T	tolerated	1	Cap3 cat B	C	*CYMP*
rs339852659	4	109671858	0.455	Agg : 0.625, Norm : 0.160,0.180,	A	G	deleterious - low confidence	0	Cap3 cat B	C	*CYMP*
rs342388220	4	111042820	0.525	Agg : 0.125, Norm : 0.560,0.740,	A	G	tolerated	0.96	Cap 1 cat B	C	*WDR47*
rs332510157	4	111098594	0.5	Agg : 1.00, Norm : 0.620,0.380,	G	T	tolerated	0.46, 051, 0.52	Cap 1 cat B	C	*CLCC1*
rs321660613	4	111116353	0.274(A), 0.42(B)	Agg : 1.00(B),0.708(C), Norm : 0.460,0.700,	G	C	tolerated	0.85,1	Cap 1 cat A,B	C	*GPSM2*
rs342779307	5	5742422	0.292	Agg : 0.344,0.292,0.375,0.321, Norm : 0.568,0.682,	C	T	tolerated	0.17, 0.05, 0.13	Cap2 cat A	D	*MCAT*
rs330946522	5	5751299	0.587	Agg : 0.083, Norm : 0.760,0.580,	T	C	NA	NA	Cap2 cat B	B	*MCAT*
rs340611738	6	15036652	0.4	Agg : 0.500, Norm : 0.880,0.920,	T	G	tolerated	0.56	Cap3 cat C	B	*PMFBP1*
rs333510524	6	15036654	0.4	Agg : 0.500, Norm : 0.880,0.920,	A	G	tolerated	1	Cap3 cat C	B	*PMFBP1*
rs333013901	9	67604156	0.2785	Agg : 0.500,0.417, Norm : 0.220,0.140,	C	G	deleterious	0.01	Cap3 cat A	C	Unknown (*PCNP*)
rs339509635	9	93078330	0.455	Agg : 0.625, Norm : 0.160,0.180,	C	T	tolerated, deleterious	0.16, 0.03,0.02,0.04	Cap3 cat C	B	*ABCB1*
rs332900783	9	117216372	0.29	Agg : 0.500,0.500, Norm : 0.260,0.160,	T	C	tolerated	1	Cap2 cat A	C	*TNN*
rs332900783	9	117216372	0.29	Agg : 0.500,0.500, Norm : 0.260,0.160,	T	C	tolerated	1	Cap3 cat A	B	*TNN*
rs319385750	10	14222171	0.41	Agg : 0.250, Norm : 0.760,0.560,	T	C	tolerated	0.96	Cap2 cat C	B	*PARP1*

Table 3.37: Part 1 of the table of SNPs selected using the variable threshold and having missense consequence. CHR and POS are the chromosome and position of the SNP in version 11 of the genome. The average.diff is the difference in allele frequencies between the aggressive and control pool, specifying the category in bracket if required. REF and ALT are the reference and alternative allele. Sift_prediction is the SIFT based prediction of the impact of the SNP, sift_score the score associated with it. Origin give the capture set and category for which the SNP was found and line the line in which it was found.

ID	CHR	POS	Average.diff	Allele.Frequency	REF	ALT	sift_prediction	sift_score	Origin	Line	Gene
rs340769264	10	14222231	0.4	Agg : 0.250, Norm : 0.720,0.580,	T	C	tolerated	0.32	Cap2 cat C	B	*PARP1*
rs328408424	10	19989785	0.365	Agg : 0.375, Norm : 0.760,0.720,	T	G	tolerated	1, 0.93	Cap2 cat B	H	*ASPM*
rs340828634	10	19997880	0.365	Agg : 0.375, Norm : 0.740,0.740,	G	C	tolerated, deleterious	0.04, 0.05, 0.06	Cap2 cat B	H	*ASPM*
rs321307993	12	40037442	0.405	Agg : 0.625, Norm : 0.340,0.100,	G	A	tolerated	0.56, 0.59	Cap3 cat C	B	*NLE1*
rs334711032	12	50073425	0.266	Agg : 0.875,0.917, Norm : 0.600,0.660,	T	G	tolerated	0.38	Cap3 cat A	B	*ZZEF1*
rs334711032	12	50073425	0.408	Agg : 0.292, Norm : 0.660,0.740,	T	G	tolerated	0.38	Cap3 cat C	C	*ZZEF1*
rs81438217	12	59139527	0.285	Agg : 0.500(C),0.750(B), Norm : 0.400,0.280,	G	C	tolerated	0.52, 0.68	Cap 1 cat A, B	B	*TRPV2*
rs325360794	12	59371738	0.3075	Agg : 0.750(B),0.625, Norm : 0.340,0.420,	C	T	NA	NA	Cap 1 cat A,B	C	*NCOR1*
rs81438358	12	59376875	0.28	Agg : 0.750,0.750, Norm : 0.340,0.600,	T	C		NA	Cap 1 cat A	C	*NCOR1*
rs692365780	12	59418694	0.3935	Agg : 0.719,1.00,0.792, Norm : 0.364,0.523,	A	G	tolerated - low confidence	1	Cap 1 cat C	D	*TTC19*
rs346399941	13	198251410	0.47	Agg : 0.750, Norm : 0.340,0.220,	A	C	tolerated - low confidence	0.62, 0.89	Cap2 cat B	C	*CLIC6*
rs337655527	13	203339714	0.37	Agg : 1.00, Norm : 0.660,0.600,	A	G	tolerated	0.1	Cap2 cat B	C	*IGSF5*
rs328154476	14	49484969	0.36	Agg : 1.00, Norm : 0.560,0.720,	A	G	tolerated - low confidence	0.14	Cap2 cat B	C	*ADORA2A*
rs81209170	16	69125456	0.306	Agg : 0.875,0.917, Norm : 0.500,0.680,	A	G	tolerated	0.23, 0.22	Cap2 cat A	C	*FAM114A2*
rs81209170	16	69125456	0.306	Agg : 0.875,0.917, Norm : 0.500,0.680,	A	G	tolerated	0.23,0.22	Cap3 cat A	B	*FAM114A2*
rs694857226	17	32583104	0.39	Agg : 1.00, Norm : 0.600,0.620,	T	C	tolerated - low confidence	0.98	Cap2 cat B	C	*AVP*
rs324827243	17	41286600	0.46	Agg : 1.00, Norm : 0.520,0.560,	C	T	tolerated	0.07,0.06	Cap3 cat B	C	*KIAA1755*

Table 3.38: Part 2 of the table of SNPs selected using the variable threshold and having missense consequence. CHR and POS are the chromosome and position of the SNP in version 11 of the genome. The average.diff is the difference in allele frequencies between the aggressive and control pool, specifying the category in bracket if required. REF and ALT are the reference and alternative allele. Sift_prediction is the SIFT based prediction of the impact of the SNP, sift_score the score associated with it. Origin give the capture set and category for which the SNP was found and line the line in which it was found

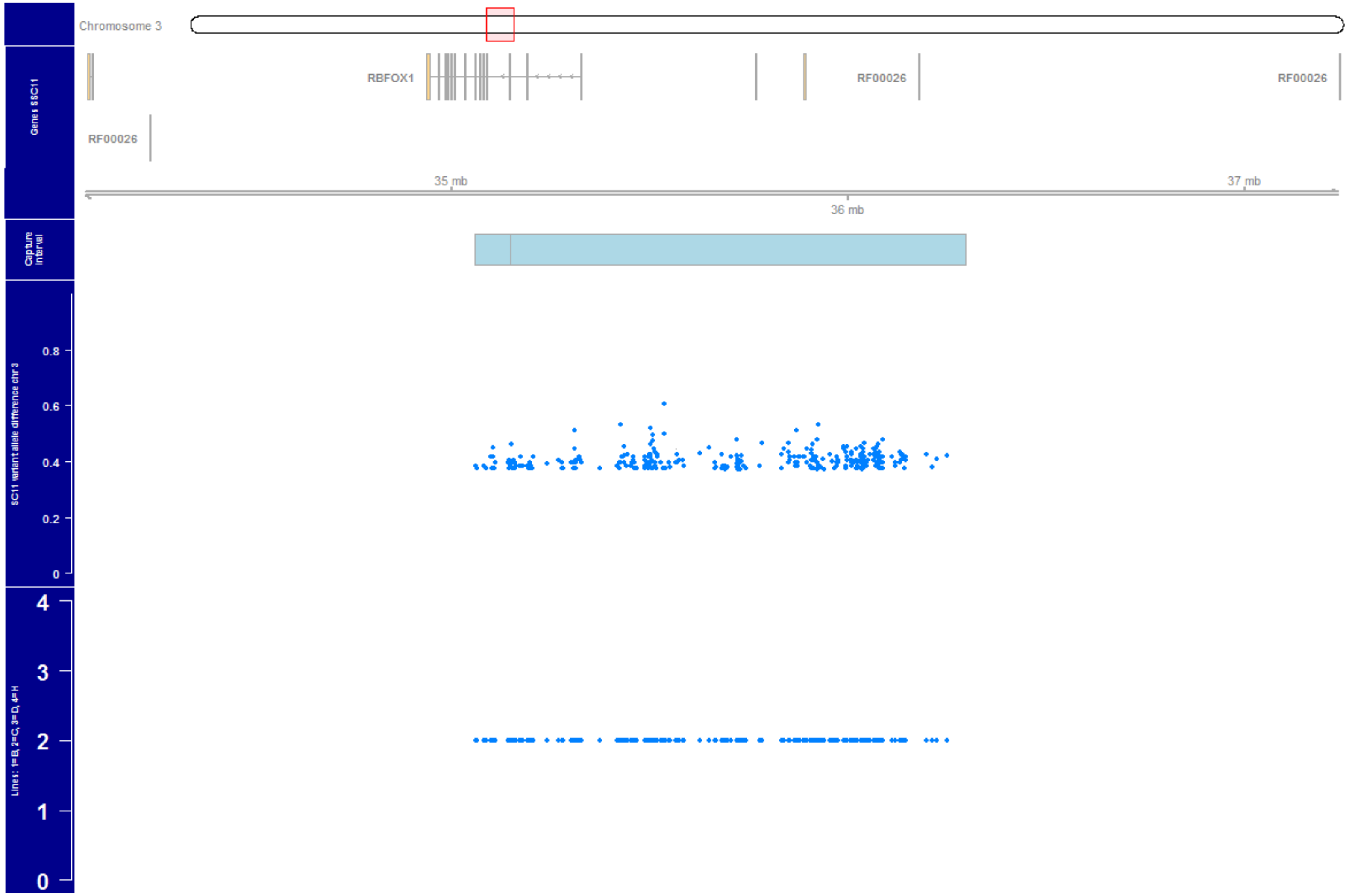


Figure 3.44: Region of interest of capture set 1 category B located on chromosome 3 at 35MB. The top of the figure show the chromosome and the region displayed (red box), below are the genes in this region and the genomic coordinates. The blue box show the region covered by the probes. Below the difference in allele frequencies between the pools (control and infanticide) are displayed for each marker above the filtering threshold. Lastly the bottom part of the graph shows the line the markers are from: 1 = line B, 2 = line C, 3 = line D and 4 = line H.

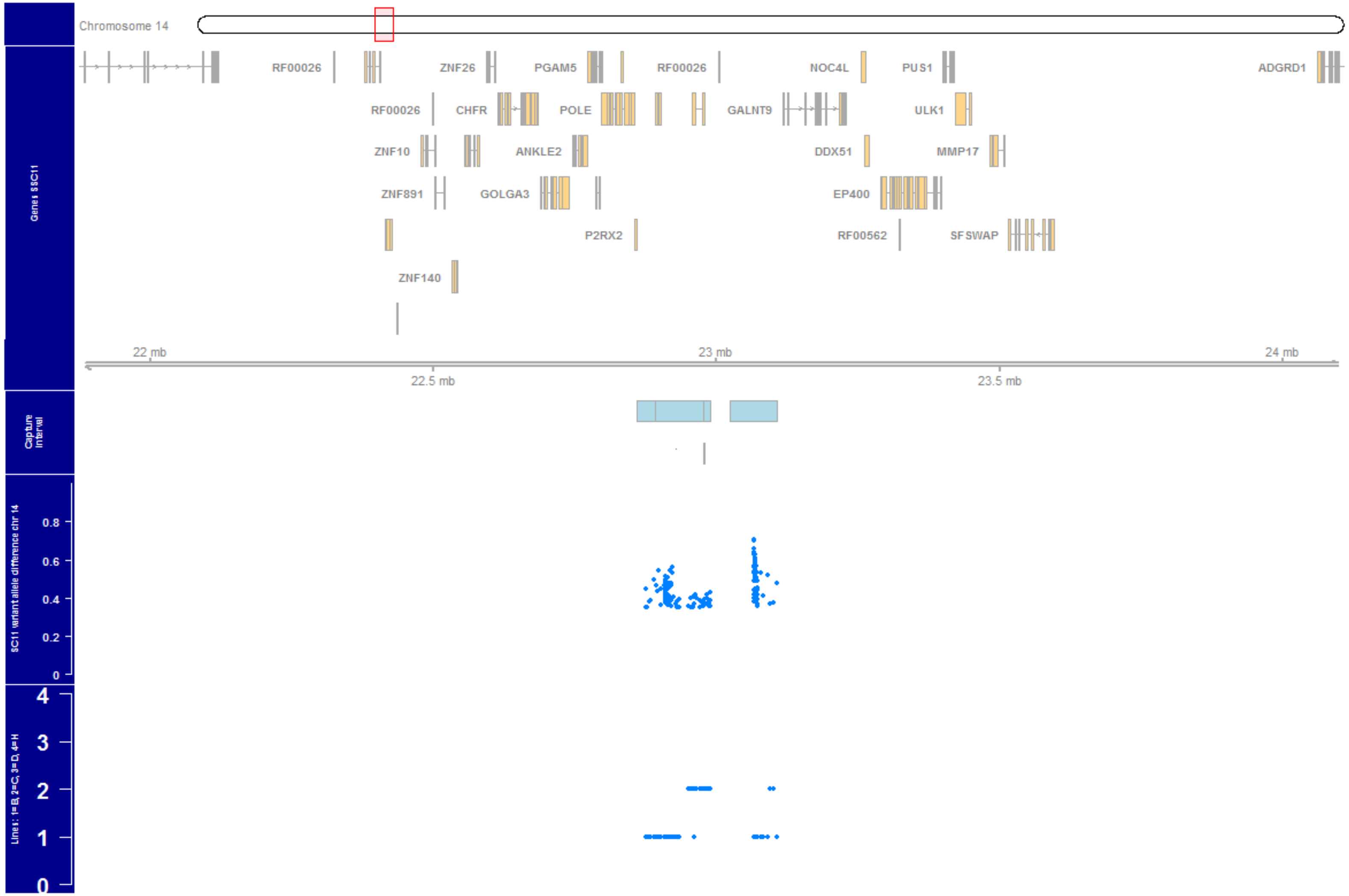


Figure 3.45: Region of interest for Capture 1 category C located on chromosome 14 at 23MB. The top of the figure show the chromosome and the region displayed (red box), below are the genes in this region and the genomic coordinates. The blue boxes show the regions covered by the probes. Below the difference in allele frequencies between the pools (control and infanticide) are displayed for each marker above the filtering threshold for that region. Lastly the bottom part of the graph shows the line the markers are from: 1 = line B, 2 = line C, 3 = line D and 4 = line H.

3.3.5 Capture 2 examples of regions of interest.

Region of interests for capture set 2 were selected in a similar fashion as for capture set 1. Figures 3.46 and 3.47 show two examples of regions of interest selected for capture set 2 category B and C. The rest of the selected region are discussed in section 4.3.6 and the genomic plots provided in supplementary material.

3.3.6 Capture set 3 regions of interest

Region of interests for capture set 3 were selected in a similar fashion as for capture set 1. Figures 3.48 and 3.49 show two examples of regions of interest selected for capture set 3 category B and C. The rest of the selected region are discussed in section 4.3.6 and the genomic plots provided in supplementary material.

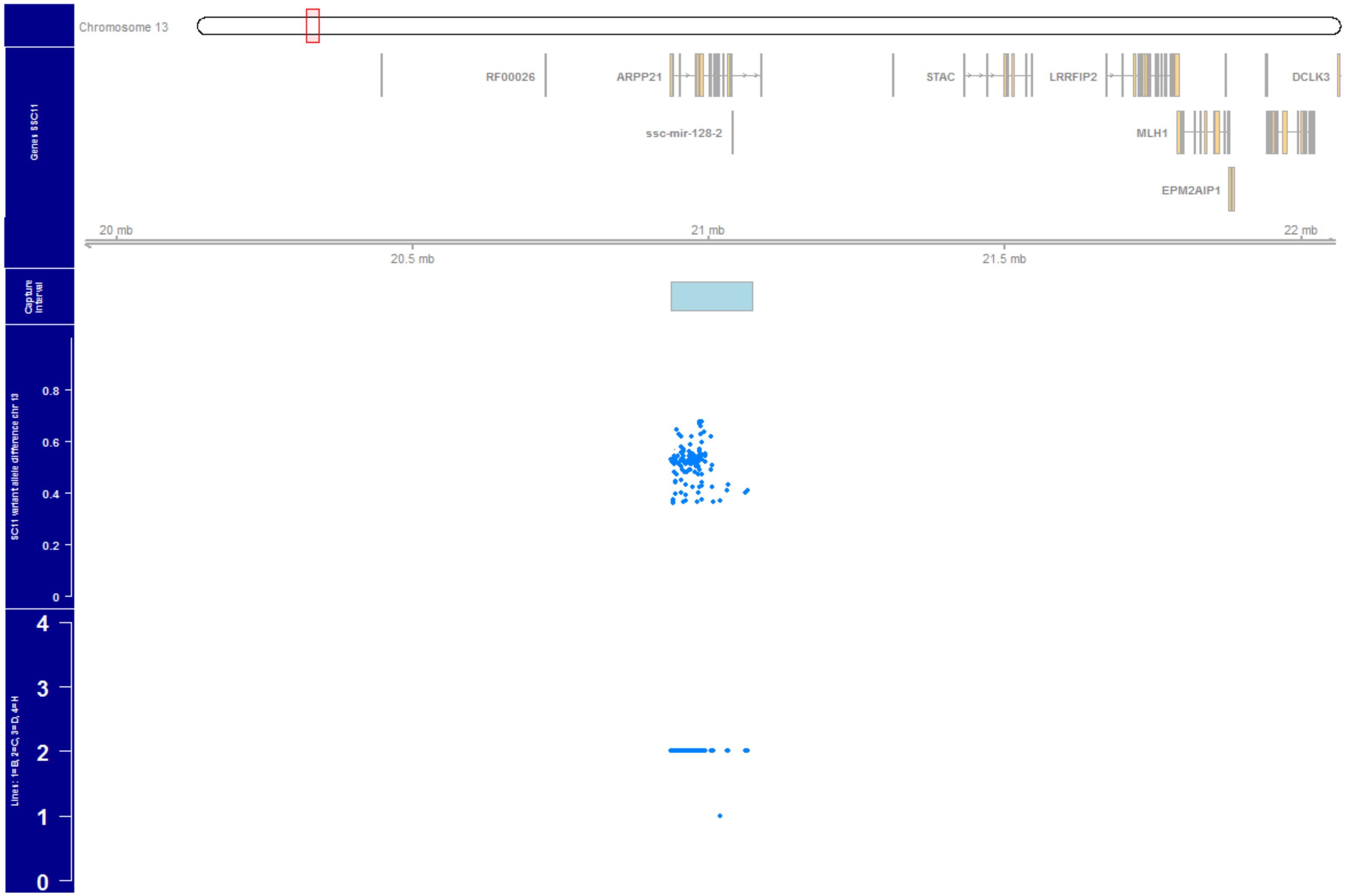


Figure 3.46: Region of interest for capture set 2 category A, covering *ARPP21*. The top of the figure show the chromosome and the region displayed (red box), below are the genes in this region and the genomic coordinates. The blue box shows the region covered by the probes. Below the difference in allele frequencies between the pools (control and infanticide) are displayed for each marker above the filtering threshold for that region. Lastly the bottom part of the graph shows the line the markers are from: 1 = line B, 2 = line C, 3 = line D and 4 = line H.

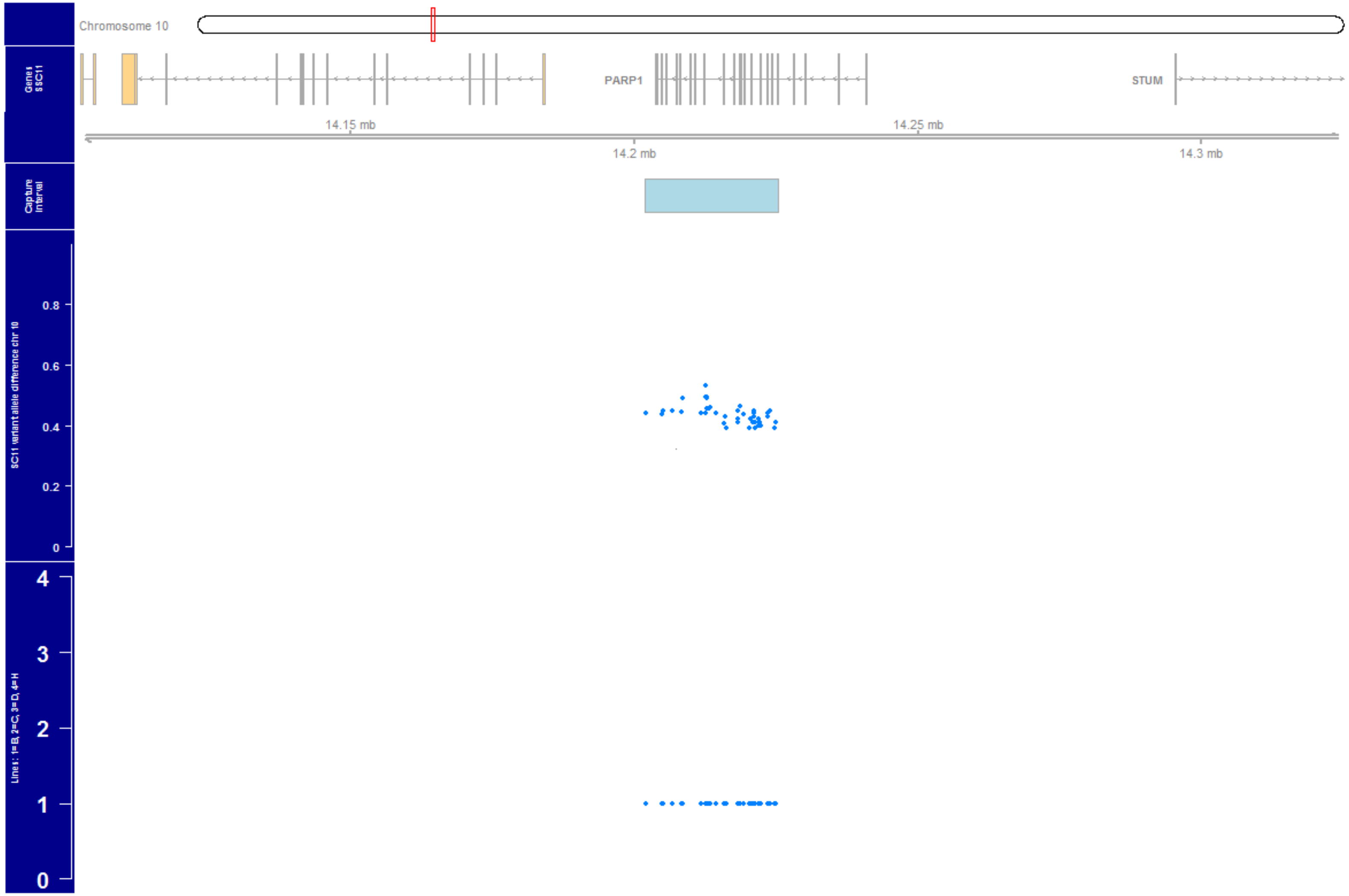


Figure 3.47: Region of interest for capture set 2 category C, covering *PARP1*. The top of the figure show the chromosome and the region displayed (red box), below are the genes in this region and the genomic coordinates. The blue box shows the region covered by the probes. Below the difference in allele frequencies between the pools (control and infanticide) are displayed for each marker above the filtering threshold for that region. Lastly the bottom part of the graph shows the line the markers are from: 1 = line B, 2 = line C, 3 = line D and 4 = line H.

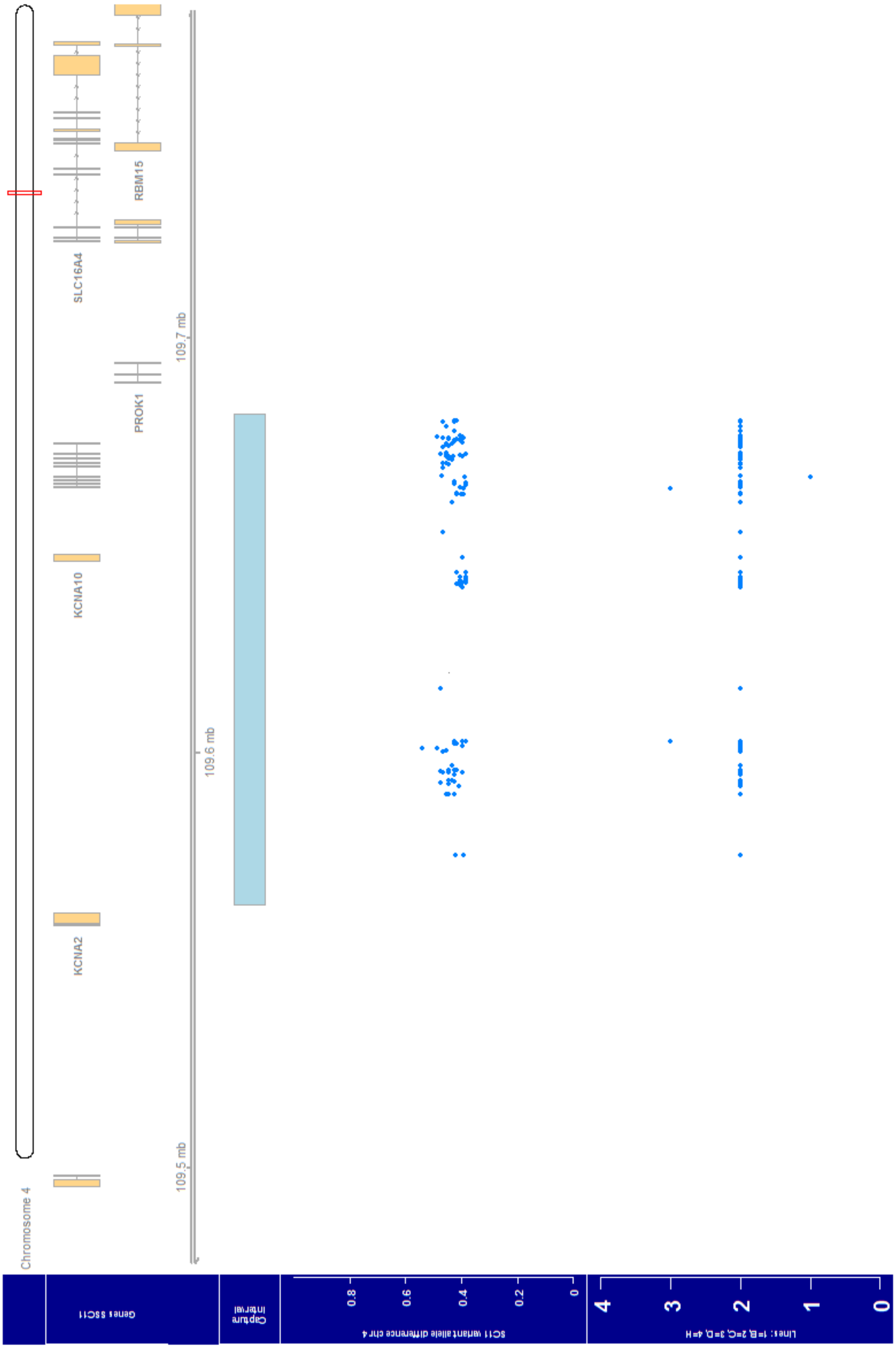


Figure 3.48: Region of interest for capture set 3 category B. The top of the figure show the chromosome and the region displayed (red box), below are the gene in this region and the coordinates. The blue box shows the region covered by the probes. Below the difference in alleles frequency between the pools (control and infanticide) is displayed for each marker above the filtering threshold for that region. Lastly the bottom part of the graph shows the line the markers are from: 1 = line B, 2 = line C, 3 = line D and 4 = line H.

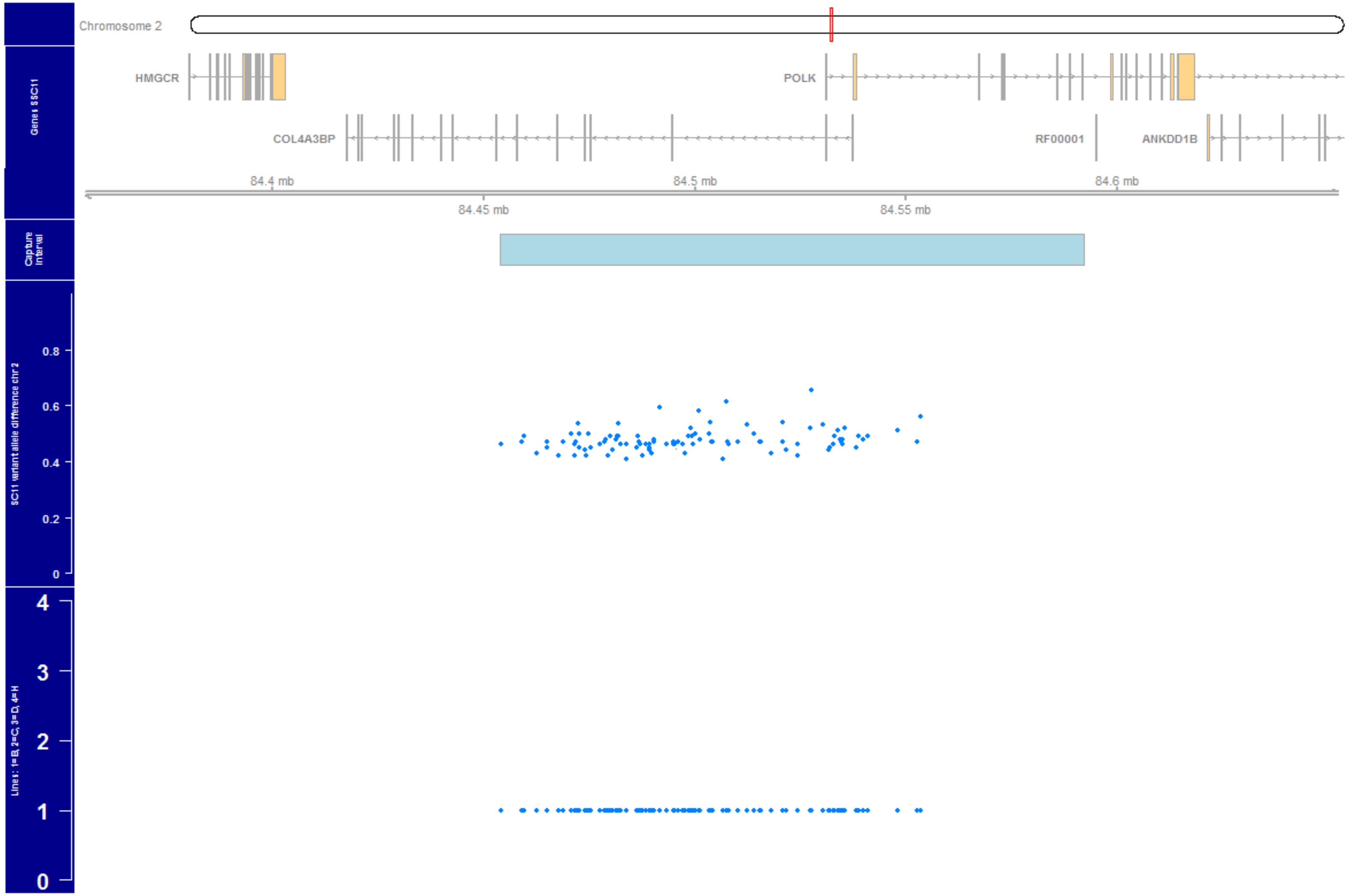


Figure 3.49: Region of interest for capture set 3 category C covering *COL4A3BP* and *POLK*. The top of the figure show the chromosome and the region displayed (red box), below are the gene in this region and the coordinates. The blue box shows the region covered by the probes. Below the difference in alleles frequency between the pools (control and infanticide) is displayed for each marker above the filtering threshold for that region. Lastly the bottom part of the graph shows the line the markers are from: 1 = line B, 2 = line C, 3 = line D and 4 = line H.

4 Discussion

The discussion focuses on the results of the different approaches that were used for this study. The first part of the discussion investigates the genotyping results from the two tests, Family Based Association Test (FBAT) and Parent of Origin (PoO), for potential targets of interest linked to the phenotype of maternal infanticide. The second part of the discussion focuses on the sequence capture sets designed using the genotyping results. The approach used to get the variants and annotate them is scrutinised and the regions of interest identified are investigated for biological functions potentially linked to the phenotype of interest.

4.1 Genotyping study

4.1.1 Genes of interest from the FBAT approach

In order to interpret the results of the FBAT analysis, SNPs showing significance in several lines were investigated for genes of interest in their vicinity. To investigate the function of the gene related to the SNP, various resources were used. First, the SNPs was called up in Ensembl [152] [184] to see if it was within a gene of interest or close to one. If no genes were found or genes of unknown functions were present, the genomic comparison tool available was used to investigate whether there was a known gene at the corresponding location in the human or mouse genome. If so, it is likely that the region in the pig covers the same gene. Once the gene was identified, the gene symbol was used to query gene cards [177, 178] and the NCBI gene database [185] in order to learn more about its function. The most interesting regions are summarised in table 4.1. Note that all SNPs that were above our filtering threshold (see section 3.1.5), were used to design the sequence capture panel.

Some of the SNPs identified by the FBAT test did reach genome wide significance and are in regions with genes linked to brain function. Some of these SNPs reach genome wide significance in several lines, such as SNPs on chromosome 2, 3, 10 and 18.

On chromosome 2, one SNP, rs81329722, reached genome wide significance in all the lines. It is an intron variant of the gene *NAV2* (neuron navigator 2). This gene plays a role in cellular growth and migration. It is also expressed in the brain in human and mouse, has been linked to age onset Alzheimer's disease in human [186], cranial nerve development in mouse [187] and in the cerebellar development and neurological disease called ataxia [188]. Ataxia is characterised by a lack of coordination in muscle movement.

The SNP on chromosome 3, rs80784123, reached genome wide significance on both lines D and H. It is intronic to the gene *TTL* which stands for tubulin tyrosine ligase. It is expressed in the brain in both human and mouse and has been linked to several psychological disorders in human, such as conduct disorder, suicide and alcohol dependency [189].

On chromosome 10, the SNP of interest, rs81325586, reached genome wide significance in all lines. It is an intronic SNP located in the *PARP1* gene. This gene codes for a chromatin associated enzyme which modifies nuclear protein by poly-ADP-ribosylation. It is involved in cell proliferation, cell differentiation and the repair of damaged DNA. It is expressed in the brain of the human and mouse. Interestingly in the mouse it has been linked to neuron healing [190], long term memory [191]. It is also involved in human in neurodegeneration, Alzheimer's and Parkinson's diseases [192, 193]. This gene has a key role in the function of neurons and their health.

The last two SNPs to reach genome wide significance in all of the lines are located on chromosome 18, rs81266760 and rs81308090. The first SNP is not located in a gene but is downstream of *SEPT7*, septin 7. There is not a clear function for the protein produced but it contains a GTP binding motif. It is expressed in the brain in both human and mouse. This gene has been linked to the *CDC42* signalling pathway which is involved in the pathology of schizophrenia and Alzheimer's disease. [194, 195]. The second SNP to reach genome wide significance on chromosome 18 is located in an intron of the gene *BMPER*, which stands for BMP binding endothelial regulator. Again the gene is expressed in the brain in mouse and human. A GWAS in human has linked this gene to Alzheimer's disease [196].

Two other chromosomes have results linked to interesting genes, chromosomes 1 and 15. Although the SNPs did not reach genome wide significance, they passed our filtering threshold in more than one line.

For chromosome 1, two SNPs are significant in two lines, lines D and H. The first SNP, rs80904152, is an intronic variant in the gene *TRPM3* which is expressed in the brain in human and mouse. The protein produced by this gene is a transient receptor potential channel which is cation-selective and functions to mediate calcium entry into the cells. Interestingly it is regulated by progesterone [197], which might indicate a role in pregnancy and potentially during labour and birth. As discussed in the previous studies section 1.4.2, progesterone blockade during late pregnancy leads to abhorrent maternal behaviours, including infanticide [58]. Furthermore, increases in levels of progesterone and oestrogen have been linked to increased levels of aggression in the pig [9]. The second SNP, rs319558321, is also significant for both lines D and H: it reaches genome wide significance for line H. This SNP is in the *TTLL11* gene (tubulin tyrosine ligase like 11). It is expressed in the brain and codes for an enzyme. Variants in this gene have been reported to influence intelligence and cognition [198, 199].

On chromosome 15, the SNP of interest, rs81278062, is an upstream variant near the gene *LRP1B* and passes our threshold for lines D and H. It codes for a low density lypoprotein receptor, which is expressed in the brain in human and mouse. This gene has been shown to be associated with Alzheimer's disease in Caribbean Hispanic individuals [200] and is also involved in cognitive decline in ageing [201].

Interestingly most of the SNPs that are found in only two lines are found in lines D and H, the lines that have the highest incidence of infanticide of the four studied.

Unfortunately none of the variants listed has any coding consequence on the gene in which they are located in or close to. While this is disappointing, it is important to keep in mind that the genome annotation of the pig is not as good as the human and mouse. These annotations are the results of an automated annotation pipeline (see section 1.7 and 4.2.5). Often the pig genome has an uncharacterised gene at a SNP location, but when comparing the region with the human and mouse genome, it becomes clear what that gene might be. As long as the pig genome is not well curated it will be difficult to assess the impact variants will have on it. It can be expected that the SNP consequences might also be inaccurately predicted due to the state of the gene annotations on the pig genome. Furthermore, the selection of SNPs for the construction of the array was not done to select SNPs that are in genes but markers that will help type LD blocks, also called "tag" SNPs. As such it is expected that a lot of markers will not map in genes. The sequence capture set was designed to target gene regions close to our candidate SNPs.

Finally, caution is required when drawing conclusions from this set of results as our threshold for selection is below the accepted genome wide significance. Despite this, there is a subset of SNPs reaching genome wide significance in several lines, and some of these genes have interesting functions and have been associated with pathology in the brain. This results were investigated further using another technology, sequence capture. The selection of the regions and the sequencing process is detailed in sections 2.8, 2.9.1 and 2.10.

4.1.2 Parent of origin genes and region of interest

The annotation method for these results is the same as used for the FBAT results (see section 4.1.1).

A summary of some of the SNPs and regions of interest is given in table 4.2. Some of the SNPs are common to the ones found in the FBAT analysis, including those that reached genome wide significance in the FBAT study. The SNP on chromosome 3, intronic to the gene *TTL* reached genome wide significance for the FBAT test. For the PO test, it was found to pass our threshold on the same tow lines: D and H. As discussed in section 4.1.1, this gene is involved in several psychological pathologies such as conduct disorder, suicide and alcohol dependency [189]. Two more SNPs also reached genome wide significance in the FBAT analysis and passed our threshold in the PO analysis. They map to chromosome 7 and chromosome 18. The first SNP, rs80949107, is an intergenic variant located downstream of the *NRN1* gene (neuritin 1). It encodes a member of the neuritin family which is expressed in differentiating neurons during the development of the central nervous system. It is

Marker	RS Identity	Chromosome	Position (MB)	Informative families	p-value	$-log_{10}(p-value)$	line	Variant type	Gene
DRGA0002016	rs80904152	1	250.3	22	1.50E-06	5.82	D	Intron variant	TRPM3
DRGA0002016	rs80904152	1	250.32	20	2.73E-06	5.56	H	Intron variant	TRPM3
INRA0007489	rs319558321	1	294.5	19	1.50E-06	5.82	D	Intron variant	TTLL11
INRA0007489	rs319558321	1	294.52	24	7.24E-08	7.14	H	Intron variant	TTLL11
ALGA0102837	rs81329722	2	42.4	71	4.76E-18	17.32	H	Intron variant	NAV2
ALGA0102837	rs81329722	2	42.4	40	3.35E-11	10.47	D	Intron variant	NAV2
ALGA0102837	rs81329722	2	42.4	36	5.59E-11	10.25	C	Intron variant	NAV2
ALGA0102837	rs81329722	2	43.3	43	5.60E-11	10.25	B	Intron variant	NAV2
ALGA0018634	rs80784123	3	45.59	39	5.62E-08	7.25	H	Intron variant	TTL
ALGA0018634	rs80784123	3	45.6	29	2.93E-09	8.53	D	Intron variant	TTL
ALGA0118632	rs81325586	10	16.6	45	4.34E-12	11.36	H	Intron variant	PARP1
ALGA0118632	rs81325586	10	16.6	40	3.35E-11	10.47	D	Intron variant	PARP1
ALGA0118632	rs81325586	10	16.6	25	5.73E-07	6.24	B	Intron variant	PARP1
ALGA0118632	rs81325586	10	16.6	30	1.21E-09	8.91	C	Intron variant	PARP1
MARC0101508	rs81278062	15	11.9	27	3.19E-07	6.5	H	Upstream variant	LRP1B
MARC0101508	rs81278062	15	11.9	25	1.14E-07	6.94	D	Upstream variant	LRP1B
MARC0083113	rs81266760	18	41.86	71	4.76E-18	17.32	H	Intergenic	Downstream of SEPT7
MARC0083113	rs81266760	18	41.9	40	3.35E-11	10.47	D	Intergenic	Downstream of SEPT7
MARC0083113	rs81266760	18	41.9	42	9.32E-11	10.03	B	Intergenic	Downstream of SEPT7
MARC0083113	rs81266760	18	41.9	36	5.59E-11	10.25	C	Intergenic	Downstream of SEPT7
ASGA0089892	rs81308090	18	42.4	36	5.59E-11	10.25	C	Intron variant	BMPER
ASGA0089892	rs81308090	18	43.2	40	3.35E-11	10.47	D	Intron variant	BMPER
ASGA0089892	rs81308090	18	43.2	43	5.60E-11	10.25	B	Intron variant	BMPER
ASGA0089892	rs81308090	18	43.23	69	1.31E-17	16.88	H	Intron variant	BMPER

Table 4.1: SNPs and regions of interest for the FBAT results

expressed in the brain in human and mouse. GWAS studies have identified various variants in this gene relating to schizophrenia, cognitive function [202, 203] and bipolar disorder [204]. The second SNP on chromosome 18, rs81233198, is also intergenic. It is located upstream of the gene *PTPRZ1,* which is exclusively expressed in the brain and encodes a member of the receptor protein tyrosine phosphatase family. It might be involved in the regulation of specific developmental processes in the central nervous system. Interestingly this gene has been linked to schizophrenia in human [205].

Some other SNPs are in common with the FBAT results but did not reach genome wide significance. There is one SNP, rs80904152, in common on chromosome 1 and mapping in the gene *TRPM3.* Another SNP, rs343881480, is in the same region and passed our threshold, but was not significant in the FBAT analysis. This SNP is intergenic according to its position in the pig genome but comparisons with the human and mouse genomes put it in the *TRPM3* gene, as well. As discussed before (see section 4.1.1), *TRPM3* is not well characterised but it is regulated by progesterone which means it could be involved in pregnancy and birth [197] and progesterone levels and has been linked to infanticide [58] and aggression in the pig [9]. Another SNP that is significant in the FBAT results and passed our threshold for the PO, rs81278062, is located on chromosome 15 and in the *LRP1B* gene. As discussed in the section about the FBAT results (4.1.1), this gene has been shown to be linked to Alzheimer Disease and cognitive decline in ageing [200, 201].

Other SNPs are specific to the PO analysis and map onto or are close to interesting genes. Two SNPs on chromosome 3, rs81373475 and rs81374014, are very close to each other and are intron variants of the gene *AUTS2.* This gene has been linked to neurodevelopment and is expressed in the brain in human and mouse. It has been linked by association to schizophrenia [206], and exonic deletions in this gene resulted in developmental delay and intellectual disability in Chinese patients [207]. In the mouse, it has correlated with impaired emotional control and cognitive memory [208], and has been shown to be expressed in regions of the brain involved in autism [209]. Another interesting SNP, rs81334603, is found on chromosome 6 also passed our threshold for the FBAT analysis, but did not reach genome wide significance. It is an intergenic variant close to the *KCNQ4* gene, which encodes for a potassium channel which plays a critical role in neuronal excitability. It is expressed in the brain in human and mouse and has been linked to depression like behaviour in mice [210]. A final SNP, rs80833324, is returned in both the FBAT and PO analyses is located on chromosome 7 and is an upstream variant of the gene *PLEKHG3.* This gene is not well characterised but it is expressed in the brain in both human and mouse, it also a candidate gene for autism risk [211]. Two other interesting SNPs that passed the threshold only for the PO analysis are found on chromosome 16 and are relatively close to each other, rs81461904 and rs81306856. The first one is an intron variant of the *MFAP3* gene, which is expressed in the brain but very little is known about its function. The second SNP is also an intron variant but of the gene *GRIA1*, which encodes for a glutamate receptor. These receptors are the predominant excitatory neurotransmitter receptors in the mammalian brain. It is expressed in the brain for both human and mouse. Several studies have linked this gene to depression and schizophrenia. One study in a Korean population [212] has found that this gene was associated with a susceptibility to schizophrenia. Another study [213] looked at the linkage of this gene in families with patients suffering from bipolar disorder. Several SNPs were linked in these families to *GRA1* suggesting a potential influence of this gene on the disease.

Overall the PO analysis did confirm five of our candidates discovered with the FBAT analysis. Six other interesting SNPs were also identified, which is not surprising given that the analysis method is different from the FBAT approach. However given the relatively lax threshold, these results need to be confirmed. This will be done by sequencing of the region of interest using sequence capture as detailed in sections 2.8, 2.9.1 and 2.10.

SNP id	CHR	SSC10 Map Coordinate	T.U_MAT	CHISQ_MAT	P_MAT	line	SNP type	Gene
rs80904152	1	250,323,029	00:22	22	2.73E-06	H	Intron variant	TRPM3
rs343881480	1	250,635,066	00:11	11	0.0009111	B	Intergenic variant	TRPM3
rs81373475	3	14,134,797	10:00	10	0.001565	H	Intron variant	AUTS2
rs81374014	3	14,332,984	11:00	11	0.0009111	C	Intron variant	AUTS2
rs80784123	3	45,593,990	1.5:23.5	19.36	1.08E-05	D	Intron variant	TTL
rs80784123	3	45,593,990	03:24	16.33	5.31E-05	H	Intron variant	TTL
rs81334603	6	157,353,343	00:21	21	4.59E-06	H	Intergenic variant	KCQN4
rs80949107	7	3,813,663	1.5:12.5	10.08	0.001496	D	Intergenic variant	NRN1
rs80833324	7	95,143,760	3.5:31.5	22.4	2.21E-06	H	Upstream gene variant	PLEKHG3
rs81278062	15	11,904,651	00:16	16	6.33E-05	D	Upstream gene variant	LRP1B
rs81278062	15	11,904,651	02:26	20.57	5.75E-06	H	Upstream gene variant	LRP1B
rs81461904	16	75,024,639	15:02	9.941	0.001616	D	Intron variant	MFAP3
rs81306856	16	75,201,733	1.5:13.5	9.6	0.001946	H	Intron variant	GRIA1
rs81233198	18	26,878,437	00:11	11	0.0009111	D	Intergenic variant	PTPRZ1

Table 4.2: Parent of origin regions of interest. CHR is the pig chromosome for the SNP, T.U_MAT is transmitted over untransmitted allele, CHISQ_MAT is the Chi squared score for the transmission of this allele and P_MAT is the associated p-value.

4.2 Sequence capture data

In this section, the quality and quantity of the sequencing data generated will be scrutinized and discussed. The efficiency of the capture set will be assessed and the latest release of the pig genome will be discussed within the context of this study.

4.2.1 Selection of the animals

The animal selection proved to be challenging for several reasons. First, the large number of animals available for selection in each line meant that manual selection using the information from the pedigree files would be very long and difficult. The initial approach was to write a script in order to select families with certain criteria but the lack of visualisation proved problematic. It was difficult to see the relationship between animals due to the complexity of some of the pedigrees. What was needed was a way of plotting the various pedigrees. Most pedigree software are relatively simple and will not allow plotting of more than a single parameter linked to individual family members. Another problem was due to the fact that some sires were shared between families, which posed difficulties for some of the software packages. After some research Madeline 2.0 [173] was identified as it could plot the families in the way needed. It allowed plotting of several parameters for each animal. Once plotted, each family had to be curated manually for individuals matching our selection criteria, see section 2.9.1. This was a lengthy process but made easier as all the information was present on the pedigree plots.

4.2.2 Read quality and pre-mapping trimming

For all three capture sets the quality of the reads is very high (figure 3.18), most reads having an average quality score above Q30 on both sequenced reads. The quality for read two can often be lower than the read quality for read one, but for this data set it is very similar: read two quality is slightly lower but the difference is marginal (see figure 3.19). Furthermore the sequencing length does not appear to have an impact on the read quality. Despite using a longer sequencing length of 100 base pairs (bp), compared to the standard 75bp, the extra 25bp do not have a large impact on the data quality.

An usual, pre-processing, the step prior to read mapping is to perform read trimming, which removes any remaining adapter sequences and will trim low quality bases from the ends of the reads. This step is usually necessary to avoid adapter sequences interfering with the mapping, or low quality bases compromising the mapping of the reads to the reference genome. The biggest impact of non trimmed reads is on the mapping efficiency which tends to drop due to the presence of adapter sequences and/or low quality bases in the read. However there is also a risk of trimming bases that are partially matching the adapters. By default most trimming software will remove a single matching base from the end of the read. Therefore a potential loss of information is possible. Quality trimming can be beneficial for the mapping, especially when the quality score drops towards the end the read, which can cause mismatches when erroneous bases are called. Our data is of very good quality, only a small proportion of reads have their quality score dropping below Q20 towards the end of the reads, and this holds for both read 1 and read 2 (See figure 3.19). The small drop in quality is not high enough to justify trimming. This avoids the removal of accurate bases that have been called with a lower quality score. Given the depth of reads obtained and the relatively small regions targetted, the coverage (see section 4.2.4) is high enough to compensate for the slight drop in quality towards the end of the reads. Furthermore, low quality score alignments are assessed when variants are called. Poor quality mapping should not result in miss-called variants in the set. Another potential issue caused by low quality bases is a drop in read complexity which might result in reads mapping to multiple locations in the genome. Read complexity represents the diversity of the bases making up the read, a longer read being more diverse than a smaller one. By trimming the low quality bases, a read can become small enough to map to multiple sites in the genome. Erroneous base calls can have a similar effect depending on the mapping quality. The read length (also called sequencing length) of 100bp should help alleviate some of these issues as the reads will be inherently more complex than 75bp reads. Therefore it should be easier to avoid the multiple mapping issue. Finally, the user guide for picard tools, which is used to prepare the aligned reads for GATK, advises not to trim reads because low quality bases are useful to the base quality score recalibration. It was decided to skip the trimming step prior to the mapping of the read on the reference genome for this project.

4.2.3 Sequence mapping

The mapping of the read sequences to the reference genome was the longest computing step of the whole analysis, a total of 60 pools needed to be mapped using a single node. It took several weeks to map all the samples as only three or four samples (pools) could be mapped in parallel.

4.2.3.1 Mapping efficiency

The mapping efficiency is high. For reads passing the Illumina filter (also known as a chastity filter), we have a 97 to 98% average of aligned reads. The chastity of a cluster is defined by the ratio between the brightest base and the sum of the brightest and the second brightest bases intensity. Clusters pass the filter if this ratio is not below 0.6 for more than 1 base during the first 25 cycles [214]. Reads (clusters) removed by this filter are usually ambiguous. This might be caused by two clusters merging together, and therefore emitting mixed signals as two fragments or more are sequenced at the same time. Overall the percentage of reads aligned is higher than we would have obtained from a whole genome approach, for which the mapping efficiency is typically around 70 to 80% [215]. The efficiency for whole genome mapping is lower because more ambiguous reads are present: regions with repeats or low complexity regions result in reads being discarded because they map at several locations. Due to the short size of reads it is not possible to resolve their placement within repeated regions. Whole genome mapping uses a masked genome to avoid regions with repeat sequences. Therefore, reads from these regions will be ignored and fall into the category of "unmapped" reads. If the unmasked genome is used, the reads originating from repeat regions will map to several locations in the genome, or if the repeat structure is unknown, stack in a specific location. In either case, the reads cannot be used. In the former case the reads are discarded, in the later case the majority of them will be filtered out during deduplication. One of the benefits of using sequence capture is its selective nature and the way the capture sets are designed (repeats are usually masked for the design). The sequences captured are usually specific to the regions they target and therefore result in a better mapping efficiency.

4.2.3.2 Read numbers

Another important aspect is the number of reads obtained which will impact on the depth of coverage of the target regions. The number of reads obtained for each of the pools and each of the capture sets is high and should result in high coverage given the size of the regions targetted. The minimum number of reads passing the filter and were mapped for all capture sets and all pools is for capture set 3, line D, pool 3 with 35.7 million reads. The highest number of reads mapped is for capture 1, line C pool 1 with 152.6 million reads. The average number of mapped reads is 81.6 million for capture set 1, 72.4 million for capture set 2 and 64 million for capture set 3. The range of values for the number of mapped reads is large, from 35 million to 152 million, and correlate with the number of raw reads. This repartition is the consequence of the pooling of the libraries. Each of the sample (pool) libraries was prepared individually and pooled together before sequencing. To get a good balance of the individual libraries in the final pool, each library needs to be accurately quantified. Unfortunately most methods have a degree of uncertainty, therefore the pool of libraries is never perfectly balanced. If several sequencing runs of the same library are planned, the pooling can be adjusted between the runs, but this was not the case for our set of libraries. A lower number of reads mapped will result in a lower coverage of the regions targetted. However despite some of the pools having significantly fewer reads mapped, the coverage remains high because the regions targetted are small in relation to the number of bases covering them (see discussion in section 4.2.4).

4.2.3.3 Duplication levels

Duplication can be a problem when calling variants as it leads to biases in allele representation or the calling of erroneous variants. If a read with a replication error from a PCR duplication gets over-represented during the sequencing and this is not addressed, it could pass as a genuine variant. Duplicate reads can also bias the allele frequency estimation, so it is necessary to remove them. Duplicate reads are reads for which the starting mapping coordinates are identical, therefore they are believed to come from the same library fragment. The duplication levels in our data set are large: between 30% to 60% of reads are duplicate reads. This amounts to a large loss of information and could be problematic. However these high levels of duplication are caused by two combined factors. The first one is that the capture targets only cover a very small part of the genome: the combined size of the sequences captured for each capture set is quite small, between 7.5MB and 13MB. The second factor is that the number of reads obtained for each pool is large, as discussed in the previous section. Combining these two factors, it is expected that a large number of reads will be covering the same fragment several times. However, it is still expected that a high coverage will be achieved after deduplication. Even for the pool with the smallest number of reads, Capture set 3, line D, pool 3 at 35.7 millions reads, with only 60% of reads left after deduplication, more than 21.4 million unique reads remain to cover a capture size of 13MB. Given that the reads are paired end 100bp, this amounts to a total of almost 4.3GB to cover 13MB, giving around 330 folds (X) coverage for the bases in the region. Thus, the depth of coverage can be expected to be very high for the captured region, which will improve confidence in the variants detected.

4.2.4 Sequence capture set: capture efficiency and coverage

Our sequence capture sets were design using three different sources, the first capture set, capture set 1, was designed using candidates region found by Quilter et al [2]. The second capture set, capture set 2, was designed using the FBAT results presented in this work. Finally, the last capture set, capture set 3, was constructed using the PO results also presented in this work. The number of probes per chromosome will vary widely between the different capture sets due to the total number of regions of interest identified .Capture set 1 has a large number of probes on chromosome 3, one of the target regions is important as it is syntenic to a region on the human chromosome 16, which has been shown to be related to puerperal psychosis [90]. Of the 11426 probes in this capture set, 7094 are targetting regions on chromosome 3, representing 62% of the total number of probes. The probes on chromosome 3 are divided across several regions, the biggest targetting the region between 26MB to 30MB, a region in which a lot of potential target genes were identified in [2]. Further along chromosome 3, a smaller region is targetted at around 35MB. It covers *RBFOX1*. Two more regions further down the chromosome are located at 100Mb and 131MB. Chromosome 1 is the second largest target, represented by 1226 probes covering intervals at 17MB, 65MB, 74MB. The other chromosomes selected (4, 6, 10, 12, 13, 14, 15 and 18) have a much lower number of probes and regions targetted, varying from 31 to 713 probes with an average of 388 per chromosomes. For capture set 2 a total of 8456 probes target the regions of interest. The largest number of probes was found on chromosome 1 and chromosome 7 with 2166 and 1476 probes representing 26% and 17.4% of the total number of probes, the other chromosomes (2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,18,X) have between 3 and 832 probes with an average of 283 probes per chromosome. The last capture set, capture set 3, is probably the most balanced in regards to the number of probes per chromosome with a total of 6636 probes. Two chromosomes are more represented that the others, they are chromosomes 1 and 9 with 981 and 935 probes each, representing 15% and 14% of the total number of probes. The other chromosomes vary between 677 and 81 probes with an average of 295 probes.

The total number of probes decreases with each capture set from 11426 to 6636. Despite the decrease in the number of probes, the size of the targetted regions increased with each capture set. For capture set 1, the size of the probes varies between 200-300bp to up to 2000-3000bp with the interquartile range covering 200 to 600bp (figure 2.2). For capture set 2 (figure 2.4) the range of sizes covered is similar but the interquartile range is shifted upwards, toward 700bp. For capture set 3 the trend continues with the the interquartile range spreading from 300bp to 1000bp. Therefore despite having significantly less probes than capture set 1, capture set 2 and 3 can select larger intervals of the genome.

Interestingly capture set 1 also has the lowest efficiency of capture compare to the other two sets. The fraction of reads covering the targetted regions (table 3.22) is between 43% and 58%, while for capture set 2 the values range between is 55% and 69% (table 3.23). Finally, capture set 3 (table 3.24) is the best performer, as the fraction of reads covering targets ranges between 69% and 72%. This is a much narrower range than the other capture sets. One possible explanation is that the protocol became more familiar for the operator, and therefore the efficiency increased. All the capture library work was carried out by the same person, Kerry Harvey. However the fact that the probes are longer for capture sets 2 and 3, along with the increase in capture efficiency points towards a better probe design, resulting in a better efficiency of capture. As the design of the different capture sets were separated by significant amounts of time, typically several months, it is likely that the technology provider improved their pipeline for probe design and efficiency.

The reads not covering the regions of interest were discarded for this analysis. Their origins could be multiple. They could arise through some off target selections. In effect the baits can capture a sequence with a partial match that maps to another location of the genome. There is also a strong possibility that contamination from genomic DNA, DNA from other regions could pass through the selection step, and therefore be amplified and sequenced, resulting in the observed off target effects. Finally some target regions have genes in them and the baits could be capturing sequences from loci of families members present in different locations on the genome.

The efficiency of capture for capture set 1 could be problematic for downstream analysis as a large number of reads are lost as they do not cover the target regions. However after checking the coverage distribution for the fraction of reads covering the probes (see figures 3.29, 3.30 and 3.31), capture set 1 has the best depth of coverage of all three sets with up to 3500X coverage and a peak at 1500X. Capture 2 and 3 go up to maximum 2000X and 1400X coverage respectively, and have peaks occur at a 100X and 700X. All three capture sets achieve at least 20X for almost all of the target bases. The normalised coverage plots (figures 3.32, 3.33 and 3.34) show that the coverage is uniform, and that at least 80% of the data is covered by more that the average coverage value for each of the sets: i.e. almost all of the targetted regions have very good coverage.

In conclusion, all of the capture sets perform well and achieve relatively good capture efficiency. Thanks to the small combined sizes of the regions targetted, very deep coverage of these regions is achieved ,which makes variant calling more accurate. It will also help ensure that the alleles coming from the animals in the different pools will be well represented providing better estimates of the allele frequency for the population the pool represents.

4.2.5 *Sus scrofa* genome release 11

During the course of this work a new version of the pig genome (11) was released by Swine Genome Consortium [150]. This new assembly of the genome highly improved on the one used for the majority of the work in this thesis. To produce this new assembly long reads were used. The sequencing was done using the Pacific Bioscience (PacBio) single-molecule real-time (SMRT) sequencing which allows for very long reads (over 10 kilo bases (kb) on average with a maximum of 60kb and with a N50 of more than 20kb) but at the price of a higher error rate (11% to 15%), although this can be moderated by several passes of the same molecule. [216].

The resulting genome is a much better assembly than the previous one (see section 1.7). With 65X coverage, it consists of 20 chromosomes and 583 unplaced scaffolds whereas the previous assembly had almost 8 times more unplaced scaffolds, 4562. The total number of base pairs and the golden path length are both lower at 2,478,44,698bp and 2,501,912,388bp. Given the long reads the assembly should produce a lower number of contigs and scaffolds, and both should be longer in size. The number of scaffolds drops by more than ten fold from 9906 to 706. The scaffolds N50 is greatly improved from 576,008bp to 88,231,837bp. The L50 is reduced from 1,303 to 9 scaffolds. Therefore a smaller number of scaffolds covers larger regions of the genome which improves the accuracy of the genome build. The number of gaps between scaffolds is greatly reduced from 5323 to 93 and the total gap length is also reduced. For 10.2 the gaps length was 289,373,899bp and for 11.1 it is 29,864,64bp, almost a 10 folds difference. The release 11 of the pig genome is therefore less patchy, with fewer uncertain or uncovered areas. Regarding contigs, their numbers between the two releases dropped immensely from 243,021 to 1,118, with the N50 increased from 69,503bp to 48,231,277bp and the L50 dropped from 8,632 to 15 contigs. The contigs forming the scaffolds are much larger and the statistics reflect that. The use of a technology with longer read produced more continuous scaffolds and contigs as reflected by the increase in the N50 and the decrease in the L50.

The number of annotated genes is also slightly higher in this build with 22,452 coding genes, 3,250 non coding genes, 178 pseudo-genes and a total of 49,448 transcripts. The annotation pipeline is similar to the one described in section 1.7 but with more sequences coming from various RNA sequencing (RNASeq) data and additional long reads from the PacBio system. The increase in the number of genes identified is modest, less than a thousand, but the increase in the number of transcripts is large, with almost twenty thousand more transcripts identified.

This latest release of the pig genome is a great improvement on the previous build and is a much more detailed and accurate representation. Two individuals were used to produce this build, the same Duroc as for build 10.2 but also a Duroc/Landrace/Yorkshire cross. This is an improvement but we can still expect some discrepancies when aligning other breeds to this genome, one of the lines used in this thesis is a Large White. The variant annotations might also still not be accurate for other breeds. The next release will hopefully incorporate more breed data, making work on animals from a different origin easier and providing a better reference for researchers.

Unfortunately for our study, the new release happened quite late in the project. Repeating the alignment steps and variant calling at this stage would have taken months, time that was unfortunately not available. Despite this drawback, the new release was still useful, as the variant locations and annotations were updated to the latest build and used to look in more details at some of the regions identified. Given that most of the variants identified by our sequencing studies are known variants, updating their coordinates will help refine the results and pinpoint the genes these variants might influence.

The capture set probe coordinates were also updated to the latest version, in order to check the impact of the new build on the regions targetted and the genes of interest. Genomic regions were plotted for the targetted regions designed from the release 10.2 against the corresponding chromosome segments from 11 and against the syntenic region from the human (details of the annotation pipeline is given in section 2.11.3 of the material and methods). Figure 4.1 show an example of one region of particular interest, the *RBFOX1* region, on the three genomes used for annotation, *Sus scrofa* 10.2 (SS 10.2), *Sus scrofa* 11 (SS 11) and Human 38 (HG38). The coordinates of the regions between SS 10.2 and SS 11 are shifted by 1MB but on SS 11 we can see that the gene of interest (*RBFOX1*) is covered partially while the locus registered on SS 10.2 is uncharacterised. This uncharacterised gene could be RBFOX1 as a search for this gene in the 10.2 build does not return any hits, but the comparison with the syntenic human genome region shows *RBFOX1* in that region. This highlight another issue with the annotation of the pig genome. Being mostly automated, with no manual curation, the annotation might not always be accurate. For example in figure 4.1 we can see that RBFOX1 is a large gene in human HG38 whereas it is relatively small in SS 11. Given that the human and the pig genome are very close genetically [217] and that the annotations used to highlight the corresponding human region rely on synteny and alignment between the two genomes, it is likely that the gene structure will be similar. In conclusion, the pig annotations might not be very accurate, but using the corresponding human region might give us more insight about the structure of a gene of interest.

Ideally the best outcome would have been to remap all the samples to the new build of the genome. This would be especially true for a different sequencing approach than sequence capture. For a whole genome approach it would have been essential to remap all the reads, as a lot of unmapped reads might be aligned to the new build. For a targetted approach, remapping of all the reads is not a necessity. If the reads mapped successfully based on the capture design of SS 10.2, it means that the design is sound as the sequences were successfully captured and therefore the design would still be valid in SS 11. If a target probe or region has no reads mapping to it, the design based on SS 10.2 was probably not accurate. Another round of design for these regions based on the new build might improve their capture. In order to assess this in more detail, the coverage of individual probes was investigated for all the capture sets and pools. Tables 4.3, 4.4, 4.5 and 4.6 display the fraction of probes for different percentages of probe length covered. The percentage of lengths covered are 0% , below 30%, below 50%, below 90% and above 90%. All of the capture sets have at least 98% of their probes where at least 90% of the probe length is covered by one read. In addition, the average coverage, median coverage and minimum coverage for the probes covered for at least 90% of their length is also displayed. The average and median coverage are very high for all capture sets. The minimum coverage can drop very low (3 reads) but given the median values (always above a 1000 reads), it is safe to assume that the coverage of this set of probes is very high. Therefore it is not necessary to remap to version 11 of the genome, most of our capture regions should be well covered. The only problematic issue is if the annotation shifts significantly, the captured region might not match close to our genes of interest any more. In this case the only solution would be to redesign the capture set using the new version of the genome.

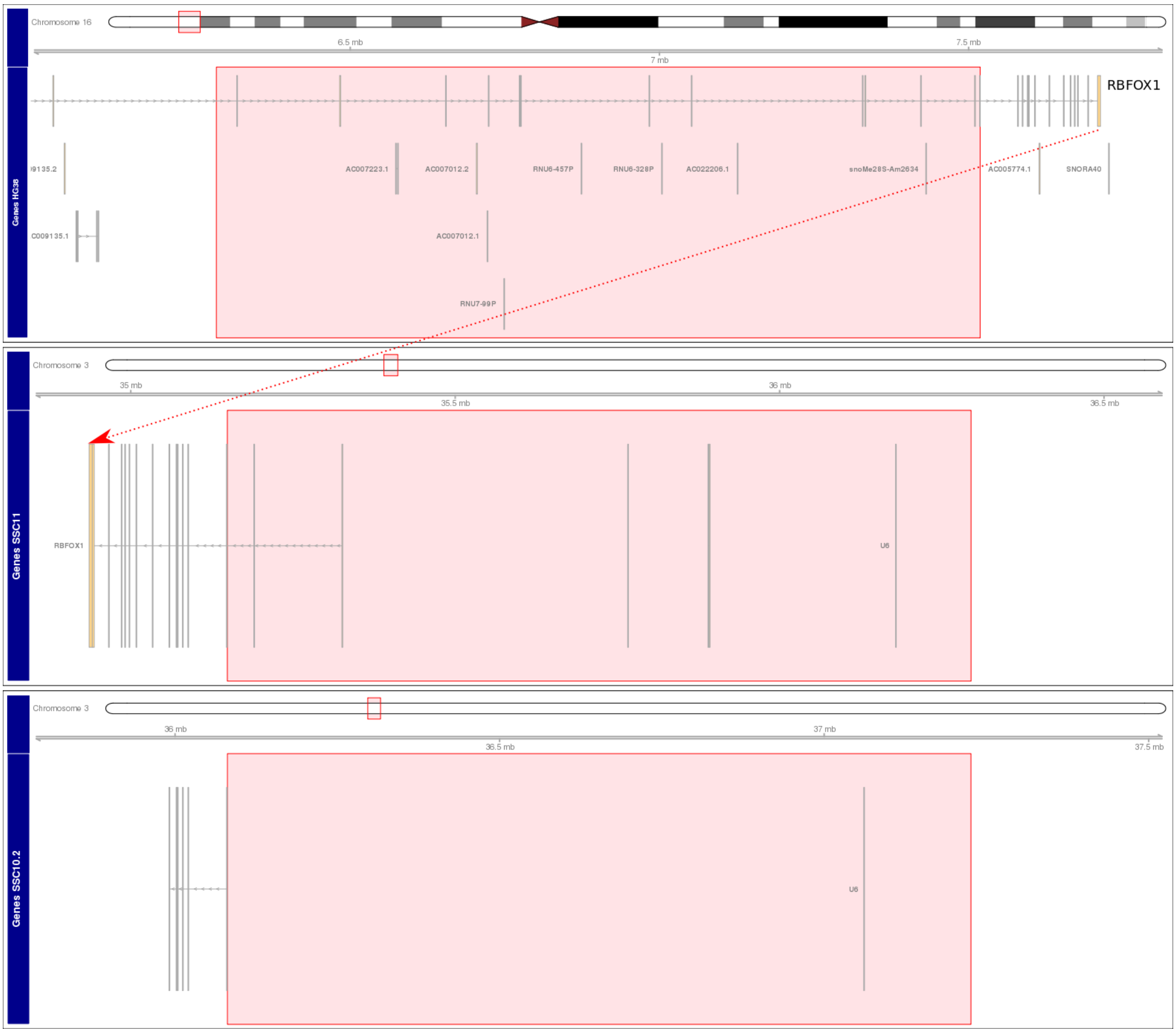


Figure 4.1: Example of capture set probe region with position in three different genomes from bottom to top: *Sus scrofa* release 10.2, *Sus scrofa* release 11 and Human release 38. Note that the gene is flipped in the pig compared to the human (red arrow).

Pools	0 coverage	below 30%	below 50%	below 90%	above 90%	average coverage above 90%	median coverage above 90%	min coverage above 90%
Line B pool 1 capture 1	0.13	0.19	0.27	0.57	99.42	6064.49	2769	7
Line B pool 2 capture 1	0.11	0.18	0.25	0.65	99.35	4563.69	2086	7
Line B pool 3 capture 1	0.10	0.16	0.22	0.56	99.44	3794.39	1741	3
Line B pool 4 capture 1	0.11	0.17	0.23	0.59	99.38	5800.68	2646	14
Line B pool 1 capture 2	0.25	0.37	0.52	1.48	98.50	5813.97	3651	7
Line B pool 2 capture 2	0.21	0.35	0.46	1.18	98.81	5272.58	3306	8
Line B pool 3 capture 2	0.20	0.31	0.41	0.88	99.11	4959.85	3135	9
Line B pool 4 capture 2	0.19	0.30	0.38	0.86	99.14	6370.22	3959	7
Line B pool 1 capture 3	0.21	0.24	0.29	0.56	99.44	5982.64	3446	7
Line B pool 2 capture 3	0.18	0.21	0.26	0.45	99.53	9356.78	5380	20
Line B pool 3 capture 3	0.20	0.23	0.27	0.51	99.49	3883.66	2232	12
Line B pool 4 capture 3	0.18	0.21	0.26	0.50	99.49	6242.23	3517	11

Table 4.3: Line B probe coverage for all capture sets. Each column displays the percentage of coverage of the probes, the percentage of the probe's sequence covered by at least one read. The average, median and minimum coverage for probe at least 90% covered is also displayed. For line B all the capture set have at least 98% of the probes covered at more than 90%

Pools	0 coverage	below 30%	below 50%	below 90%	above 90%	average coverage above 90%	median coverage above 90%	min coverage above 90%
Line C pool 1 capture 1	0.11	0.18	0.26	0.66	99.34	6185.72	2826	9
Line C pool 2 capture 1	0.12	0.18	0.26	0.73	99.26	3388.50	1567	4
Line C pool 3 capture 1	0.10	0.17	0.25	0.67	99.32	3407.91	1551	5
Line C pool 4 capture 1	0.13	0.18	0.26	0.57	99.43	4180.06	1937	4
Line C pool 1 capture 2	0.22	0.34	0.45	1.17	98.82	4753.66	2953	7
Line C pool 2 capture 2	0.21	0.38	0.50	1.27	98.73	5480.80	3440	8
Line C pool 3 capture 2	0.26	0.35	0.48	1.15	98.85	5168.04	3196	4
Line C pool 4 capture 2	0.20	0.37	0.45	1.10	98.89	5316.06	3301	5
Line C pool 1 capture 3	0.18	0.21	0.24	0.47	99.52	9956.60	5659	21
Line C pool 2 capture 3	0.21	0.24	0.29	0.59	99.41	6436.54	3662	10
Line C pool 3 capture 3	0.20	0.21	0.24	0.44	99.55	7993.41	4592	12
Line C pool 4 capture 3	0.17	0.21	0.24	0.45	99.53	7056.89	4000	6

Table 4.4: Line C probe coverage for all capture sets. Each column displays the percentage of coverage of the probes, the percentage of the probe's sequence covered by at least one read. The average, median and minimum coverage for probe at least 90% covered is also displayed. For line C all the capture set have at least 98% of the probes covered at more than 90%.

Pools	0 coverage	below 30%	below 50%	below 90%	above 90%	average coverage above 90%	median coverage above 90%	min coverage above 90%
Line D pool 1 capture 1	0.13	0.20	0.27	0.60	99.39	1933.59	892	2
Line D pool 2 capture 1	0.14	0.18	0.27	0.60	99.40	2456.37	1111	3
Line D pool 3 capture 1	0.10	0.15	0.21	0.55	99.44	2849.38	1308	6
Line D pool 4 capture 1	0.11	0.14	0.23	0.56	99.44	3089.87	1419	3
Line D pool 5 capture 1	0.10	0.18	0.25	0.60	99.39	2776.64	1253	5
Line D pool 6 capture 1	0.10	0.16	0.23	0.55	99.44	2600.52	1198	5
Line D pool 1 capture 2	0.17	0.26	0.38	0.84	99.15	5298.53	3342	12
Line D pool 2 capture 2	0.21	0.30	0.37	0.89	99.11	6306.72	3997	12
Line D pool 3 capture 2	0.18	0.27	0.38	0.86	99.12	5701.80	3599	7
Line D pool 4 capture 2	0.20	0.30	0.38	0.91	99.08	8411.30	5217	9
Line D pool 5 capture 2	0.19	0.32	0.40	0.91	99.08	5502.87	3452	6
Line D pool 6 capture 2	0.20	0.28	0.35	0.85	99.12	5714.27	3631	10
Line D pool 1 capture 3	0.20	0.21	0.27	0.47	99.53	9346.25	5361	10
Line D pool 2 capture 3	0.20	0.24	0.29	0.53	99.46	6898.77	3964	23
Line D pool 3 capture 3	0.20	0.23	0.29	0.51	99.49	5852.13	3327	6
Line D pool 4 capture 3	0.20	0.23	0.27	0.50	99.50	6141.30	3545	2
Line D pool 5 capture 3	0.21	0.23	0.27	0.47	99.53	7335.80	4137	12
Line D pool 6 capture 3	0.18	0.23	0.29	0.51	99.49	5769.24	3239	8

Table 4.5: Line D probe coverage for all capture sets. Each column displays the percentage of coverage of the probes, the percentage of the probe's sequence covered by at least one read. The average, median and minimum coverage for probe at least 90% covered is also displayed. For line D all the capture set have at least 98% of the probes covered at more than 90%.

Pools	0 coverage	below 30%	below 50%	below 90%	above 90%	average coverage above 90%	median coverage above 90%	min coverage above 90%
Line H pool 1 capture 1	0.10	0.17	0.24	0.52	99.48	3814.03	1750	6
Line H pool 2 capture 1	0.11	0.14	0.22	0.52	99.48	3938.69	1769	3
Line H pool 3 capture 1	0.11	0.15	0.23	0.55	99.45	2862.29	1315	5
Line H pool 4 capture 1	0.10	0.16	0.23	0.56	99.44	2471.97	1140	5
Line H pool 5 capture 1	0.11	0.18	0.25	0.60	99.40	2726.34	1250.5	4
Line H pool 6 capture 1	0.10	0.15	0.23	0.55	99.43	3226.56	1476	8
Line H pool 1 capture 2	0.22	0.33	0.43	1.05	98.94	7322.24	4552	16
Line H pool 2 capture 2	0.25	0.34	0.46	1.12	98.84	5086.74	3154.5	7
Line H pool 3 capture 2	0.20	0.35	0.43	1.03	98.95	5081.47	3189	5
Line H pool 4 capture 2	0.18	0.28	0.35	0.83	99.17	3943.07	2454.5	7
Line H pool 5 capture 2	0.22	0.33	0.44	0.98	98.99	4754.84	2962	8
Line H pool 6 capture 2	0.20	0.31	0.38	0.82	99.17	5232.68	3236.5	7
Line H pool 1 capture 3	0.21	0.24	0.29	0.54	99.46	7709.16	4360	19
Line H pool 2 capture 3	0.20	0.24	0.29	0.53	99.46	5843.64	3238.5	20
Line H pool 3 capture 3	0.21	0.24	0.30	0.51	99.47	6913.85	3917	9
Line H pool 4 capture 3	0.20	0.23	0.26	0.53	99.46	5196.68	2930.5	7
Line H pool 5 capture 3	0.21	0.23	0.27	0.47	99.53	7035.24	4003	10
Line H pool 6 capture 3	0.20	0.23	0.27	0.47	99.53	6551.80	3654	19

Table 4.6: Line H probe coverage for all capture sets. Each column displays the percentage of coverage of the probes, the percentage of the probe's sequence covered by at least one read. The average, median and minimum coverage for probe at least 90% covered is also displayed. For line H all the capture set have at least 98% of the probes covered at more than 90%.

4.3 Genome Analysis Tool Kit (GATK) and variant processing

4.3.1 Post mapping read processing

Once read mapping is completed, the aligned reads have to be processed through the GATK pre-processing steps in order to make sure that the input data used for the variant calling is of good quality. The deduplication step removed a lot of reads, but large numbers are still left for each of our pools, which should give us enough depth for accurate variant calling. The depth of sequencing must be high enough to unsure that every animal in the pool is sufficiently represented to contribute in a significant manner to the estimation of the allele frequencies within the phenotypic group it represent. As discussed previously (section 4.2.4) the coverage is very deep and therefore should not have a negative influence on the variant calling.

Once duplicate reads are removed, the next step is to realign the reads around indels. This is an important step as read mapping will be affected by the presence of structural variants (insertions and deletions, also called indels) and some reads might be discarded, or get low quality scores, because of the presence of indels. Correcting potential misalignments is done by the realignment step of the GATK pipeline, using a reference file of structural variants. Our reference file for structural variants for the pig is from dbSNP, version 145, which is based on the genome build SS 10.2. As discussed in the introduction (see section 1.7), this reference has a large number of gaps and scaffolds not placed on the genome yet. Therefore some of the structural variants will be uncharacterised or belong to scaffolds that have not been placed. The consequence of the patchwork nature of this genome is that uncharacterised indels present in our data will be unaffected by the remapping step, which might reduce the mapping quality for these locations.

The next important step is the read quality score recalibration. The mapping quality score is a phred score. It uses the same scale as the read quality discussed previously and can be used in the same way. A read quality of 30 has a one in a thousand chance of being wrong. Mapping quality consists of different source of mapping error, contamination, heuristic error (error caused by the alignment method) and error caused by the repetitiveness of the reference (see main text and supplementary text [218]). The first two components are difficult to estimate, but the last can be estimated. The error for *bwa-mem* is estimated by $10/log10*[log4*(S_1-S_2)-log(n_{sub})]$ with S_1 being the best alignment score and $S_2$the second best and $n_{sub}$the number of suboptimal alignments for a given read. According to [144], *bwa* tends to overestimate slightly the quality score. While this is not a huge problem, according to the GATK developers, the quality score is often overestimated around base deletions , insertions or substitutions. Therefore, around known locations with this type of event, GATK will adjust the quality score. For this adjustment the software needs a reference file in order to know which loci have to be adjusted. The reference used is *Sus scrofa* dbSNP 145. Similarly to the indel reference, this version of dbSNP is based on SS 10.2, which will have the same drawbacks discussed for the read realignment recalibration. Despite of these drawbacks it was still beneficial to proceed with the recalibration for bases covering known variants at the time this step was performed.

For both the realignment and base recalibration steps, the nature of the data helps alleviate some of issues raised above. Given that we are not using a whole genome approach but a targetted approach, it is less likely that our variants will fall into an unplaced scaffold or into a gap between scaffolds. If a read maps to a target region, this region should not be a region of uncertainty. As discussed in the section on release 11 of the pig genome (section 4.2.5), a very small proportion of the capture probes have no, or a small proportion of, reads covering them. Most regions for the capture sets are well covered. Therefore it is unlikely that targetted regions are problematic. If the probes targetting the region were designed on erroneous sequences, the results would be that no reads are mapping to the region. Therefore it can be assumed that targetted regions in this study are fairly well characterised since the capture was successful and the mapping to these regions was good. Structural and sequence variants should therefore be well characterised.

The quality scores before and after recalibration show that, as expected the aligner overestimated the read mapping scores around structural and sequence variants (see figures 3.35, 3.36) which is the outcome that was expected.

4.3.2 Variant calling and pooled samples

4.3.2.1 Algorithm discussion

The GATK suite has several algorithms for variant calling, the most recent one is the *HaplotypeCaller* [219]. Its strength comes from calling variants on several samples at the same time. *HaplotypeCaller* works in four steps. First it determines *active regions* using the aggregated data from the different samples. Active regions are regions with substantial variations compared to the reference, therefore excluding regions of high similarity to speed up the process. Next it determines haplotypes by building an assembly (DeBruijn graph) graph using the reference as a template. Nodes in the graph correspond to mismatches with the reference, and *HaplotypeCaller* selects the most likely graph using the weight of evidence, i.e. the number of reads supporting each graph. Once the graphs are built, the next step is to evaluate each haplotype. In order to perform this evaluation the reads are aligned against each of the possible haplotypes using PairHMM, HMM stands for Hidden Markov Models. These models use available information to infer missing information. For example if indoors with no windows, one can infer that it is raining outside if someone comes in with wet shoes. This might be wrong for several reasons but it is a reasonable assumption under normal circumstances. The PairHMM uses the base quality score for the HMM, it helps to infer if the variant is likely to be real or not. The last step is to assign the genotype back to the individual samples, which is done using a Bayesian approach to compute likelihood ratios. The strength of this approach is the combination of data from a cohort of samples and using it to call variants. The whole process is broken down into several steps. First calling the variants on individual samples and generating the variant files for each sample to then consolidate them together and do a joint cohort variant calling. This produces raw variants that can be filtered using several different methods. The advantage of the *HaplotypeCaller* method compared to previous algorithms such as the *UnifiedGenotyper* is that it is better at calling insertions and deletions, as they are part of the Hidden Markov Models whereas the *UnifiedGenotyper* does not model insertions or deletions.

Unfortunately for this study the use of *HaplotypeCaller* was not possible. The first step of the variant calling, calling variants for individual samples proved too problematic for our samples. As our samples are made from pools of several individuals, one of the parameters we had to change is the ploidy (the number of chromosomes). The smallest pool had four animals and the largest had 46, which means the ploidy varied between eight to 92 chromosomes. This resulted in problems using *HaplotypeCaller* as our computing node was running out of resources. Our initial computing node had 125 GB of RAM which was thought to be sufficient. However while running this step the node kept crashing due to a lack of memory. Upon investigation it was noticed that the crash occurred at a particular locus. The parameters of the step were modified to skip this locus but this only shifted the problem to another locus. After excluding several loci it was decided that this approach was unsustainable. Another solution was to upgrade the available resources. The node was upgraded to double its memory at 250GB. Unfortunately the software was still crashing due to a lack of memory. Another computing node was found to run this step, with 500GB of RAM, but again the software was running out of memory. At this point the GATK developers advised us to use an alternative and older algorithm available that can cope with this high level polyploidy, the *UnifiedGenotyper* [149].

The *UnifiedGenotyper* is a simpler variant caller compared to the *HaplotypeCaller* but it can handle the very large ploidy present in our samples. It is a permissive caller and therefore will call large amounts of variants, and some will be false positive. It can call variants in multiple samples at the same time but for the animal pools that were used, and because of the large ploidy in some of them, it was deemed preferable to call variants on individual pool. The *UnifiedGenotyper* performed well on all the pools and no issues with memory usage were encountered while running it on our cluster node.

4.3.2.2 Variant filtering

Because of the use of the *UnifiedGenotyper*, the choice of filtering was limited to hard filters. It was not possible to use the most advanced filtering tools based on machine learning as they require several sets of high quality known variants to train the model. Typically, for human the sets come from dbSNPs, the 1000 genomes project, genome in a bottle, etc... Unfortunately such sets are not available for the pig yet, only dbSNP has been released and its content might not be the most accurate for all of our lines, as it is based on the Duroc reference. Thus the use of hard filters was the only choice available. The values used for filtering (see section 2.10.3) are the ones recommended by the GATK team. They are standard values for filtering and should give a good balance to filter the variants called. Using these filtering parameters will avoid being too stringent which would miss potentially interesting polymorphisms.

The number of raw variants ranges between 23,000 and 57,000 per pool across all capture set (see table 3.25). Overall capture set 2 has the lowest number of raw variants called with 609,533 and capture set 3 has the largest with 976,980. After filtering, the lowest quantity of variants is still for capture set 2 with 385,016 and the largest for capture set 3 with 655,310. The filtering does remove between 30% to 40% of the variants, but the precise fraction of variants found and filtered between samples varies. For capture sets 1 and 2 the number of variants found and filtered is fairly similar. For capture set 3 however it is higher by at least 10,000. This might be a consequence of the size of the regions targetted. For capture set 1 the probes target 7MB, for capture set 2 10MB and capture set 3 13 MB. Surprisingly despite the difference in capture size between capture set 1 and 2, and 2 and 3, the number of variants found after filtering is similar for capture set 1 and 2 but much larger for capture set 3.

The number of variants found for each capture set generally correlates well with the number of probes targetting a given chromosome. The only exception to this is from capture set 1, where the number of variants found for chromosome 1 for the different pools is lower than for other chromosomes. Chromosomes 4, 12, 15 and 18 have a higher number of variants identified compared to chromosome 1 despite having around half of the number probes targetting chromosome 1.

4.3.3 Variant comparisons

As described in section 2.11.1, the variants were compared between pools in three categories. Category A compared all the infanticide pools against the control pools, category B compared the serial infanticide pools against the controls, and category C compared the pools with a family history of infanticide to the controls pools. The first category helps to identify variants and regions that are common between the serial infanticide animals and the animals with a family history of infanticide. The category B and C prospects will identify variants and regions that are more specific to the strength of the trait for the serial offenders or linked to heritability for the family history pools. In order to chose the thresholds for the comparisons, the control pools were used to checked the repartition of the difference in allele frequency (see figures 3.37, 3.38, 3.39, 3.40, 3.41, 3.42 and 3.43). The variations in allele frequencies between the two control pools for the various capture sets are fairly similar following their repartition. The number of variants with small differences in allele frequency is large and the number of variants with large differences in allele frequency is small. Most of the differences in allele frequencies are found to be below 30% (0.3).

Because most of the variability between the two control pools is segregated below 30%, the base threshold for removing highly variable variants between pools of the same category is set at 30% for both the control and the infanticide pools. Due to the lower number of individuals in some of the infanticide pools, the allele frequency variations between infanticide pools of the same category are more pronounced (see supplementary data).

To compare allele frequencies between controls and infanticide pools two approaches were taken. The first approach used hard thresholds for the comparisons between infanticide and control pools. The first threshold was set at 30%, the same as the within pool cut off. The approach worked well for category A (comparing control pools against infanticide pools) but returned a large number of variants for categories B and C. Therefore the threshold was increased to 50% allele frequency difference between the control and infanticide pools.

The second approach used a more variable threshold based on the distribution of the differences in allele frequency of the variants (see section 2.11 for details). The distribution of the allele differences between the control pools, for each capture set, and the distribution of the allele differences between infanticide and control pools, for the different capture set and category, reveal some interesting information. The allelic differences between the control pools for each line already revealed that most of the allele variation is found to be below 30%. Grouping the control pools per capture set confirmed that trend. The threshold of mean plus three standard deviation was just above the 30% allele frequency difference (see figure 3.37). When grouping the lines per capture set and category, and comparing infanticide against controls (see figure 3.37), a similar picture emerges but the allele differences are more important. For the category A pools, the thresholds of mean plus three standard deviation for each capture set is similar to the one obtained when comparing the control pools. For category B and C the thresholds are higher and the distributions have a longer tails. There are multiple potential reasons for this. First, the pool number are not balanced between the lines and categories. While the control pool have 25 animals in each of them, the infanticide pools have more variable numbers. The serial infanticide pools have a low number of individuals with 4, 6, 14 and 4 animals for line B, C, D and H respectively. The family history of infanticide pools have more animals with 4, 12, 36 and 32 animals for line B, C, D and H respectively. This imbalance in the number of animals in the pools will results in more extreme allele frequency from pools with a small numbers of animals compare to pools with a larger numbers of animals. Furthermore, for category A, we are comparing two pools against two pools which might give a better estimate of the allele differences and reduce any outlier effect. For categories B and C, the two control pools are in some case compared against a single infanticide pool. Indeed for line B and C there are only a single infanticide pool in each category, while for lines D and H, several infanticide pools are available for category C (family history of infanticide). The presence of single pools in categories when comparing the allele frequencies will results in more extreme values. Therefore the overall distribution will be influenced by these extreme values. It is also worth noting that category A groups the two infanticide sets, serial infanticide and family history which could lessen any specific effect caused by alleles linked solely to one of the two categories. The fact that the results of the selection of the variant in category C has a large proportion of variants selected on some chromosomes (see tables 3.30, 3.33 and 3.36) when using the variable thresholds, seems to suggest that these might be some category specific effects. This might not be visible for the hard threshold approach as it might be too stringent for the effect to appear.

The control pools were also used for a mock comparison, to compare the proportion of selected variants with the comparisons against the infanticide pools (see tables 3.26 and 3.27). The percentage of selected variants in the mock comparison is lower by one to two percent for capture sets 1 and 2, but are similar for capture set 3. The comparisons with the infanticide pools are selecting more variants than the comparisons between two similar pools.

The thresholds based on the distribution of the allele frequency differences, using mean plus three standard deviations, selected a much higher number of variants. The thresholds were usually lower than the hard thresholds that were chosen before. It was decided to use this method of filtering to select variants and regions of interest. The percentage of selected variants is similar between the different capture sets and categories. Category A selecting between 3 to 3.8% of the total number of variants, between 3.6 and 4.3% for category B and between 4.7 and 6.5% for category C. It is interesting that the highest number of variants selected is for category C, which compares the control pools with the ones for individuals with an history of infanticide. Some chromosomes in this category have a very large number of variants selected. For capture 1 category C, the number of variants selected on chromosome 6 represent 50% of the total number of variants present, furthermore 32% of the polymorphisms on chromosome 13 were selected for the same capture set and category. For capture 2 category C, 42% of the variants on chromosome 11 were selected. For capture 3 category C, again we have 26% of the variants on chromosome 6 selected and 34% for chromosome 10. It is interesting that this skew in selected variants proportions are happening only for category C and not for any of the other categories. Therefore these regions might be of interest as they are clearly outliers when compared to other chromosomes and other categories.

On the other hand, some target regions did not select any variants. This is expected, all of the target regions identified using the genotyping data might not be relevant. The genotyping tests were used to identify potential regions of interest but not all were expected to be relevant after a more in depth investigation. It was not expected that a large number of variants will be of interest, the variants identified will help refining the regions of interest and investigate them in more details. Using annotation pipelines and updating the coordinate to the version 11 of the genome (see section 2.11.2 and 2.11.3), the variants selected can be investigated in more details to evaluate their consequences on the genes present in the region.

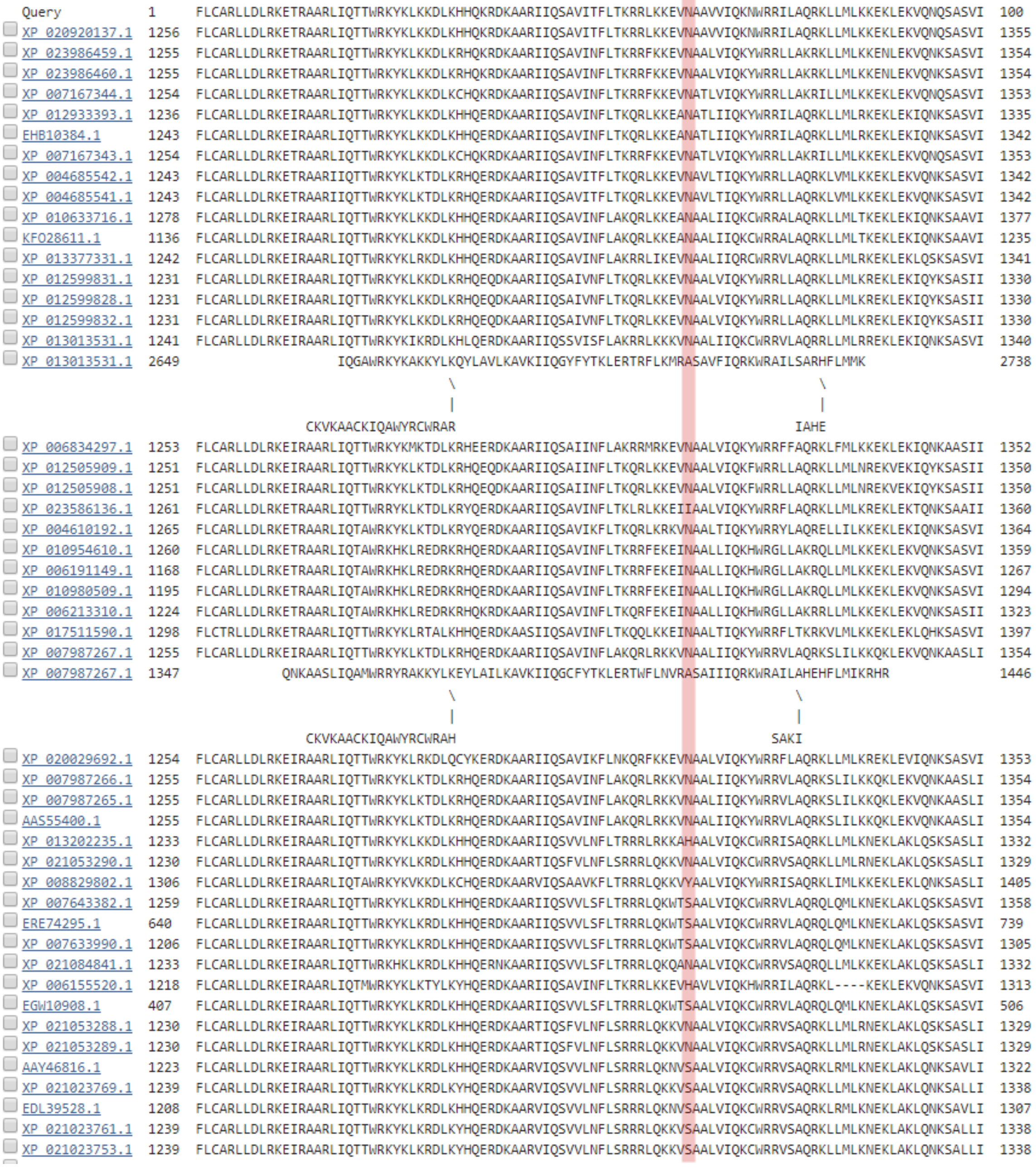
4.3.4 Variant annotation

The variants annotation has proven to be one of the most challenging steps of this work. When the first set of results was curated, only version 10.2 of the pig genome was available. The annotations were poor and making sense of the results proved to be difficult. It was decided to develop a pipeline to compare the variants to the human genome in order to see if any gene or human variant of interest was located in the corresponding region of the human genome. Thankfully, homology information and alignments between the pig and human genomes were available for this task. This allowed correspondences matching between pig and human, but sometime the alignments were poor. A more bespoke annotation pipeline was needed and the pipeline detailed in section 2.11 was designed. This pipeline was applied to the hard threshold results. It was a lengthy process, as for each gene identified, its function needed to be curated manually. While this part of the work was on going, the new version of the pig genome was released and it was decided to change the annotation pipeline in order to use this updated reference. The objective was to update the coordinates of the variants to the new version of the genome and hopefully get better gene annotations. The first attempt to update the coordinates worked relatively well but some SNPs were lost during the process. Because the data used were from the hard thresholds initially applied, there were only a small number of SNPs selected, and losing even a fraction of them could be problematic. The pipeline described in section 2.11.3 was designed in order to try to update as many variants as possible. The pipeline is complex and applying it for each capture set, line and category was a long process. Soon after completion of this step the protocol was inspected for a less arbitrary filtering method that could be applied to select the variants. The thresholds based on the distribution of the data were implemented, and instead of trying to update all the SNPs, it was decided to simply update the coordinates to version 11 of the genome. This was done by querying Ensembl biomart using the R package and the SNP identification lists to get back the coordinates in version 11 of the genome. Any SNP lost was probably not very interesting, if it was not a valid SNPs in the latest version of the genome, its interpretation would have be difficult. Plotting the region where the SNPs were located also helped for the interpretation and identification of genes. The gene function was followed up manually, as described in section 2.11.4. This means that a lot of the work that went into the interpretation and curation of the first set of variants, and the more bespoke pipelines used to update the variants, have not been applied to the final set of results. These processes can still be applied in future work or used to update variants in other species that might not yet have benefited from an improved genome.

4.3.5 Missense SNPs and their targets

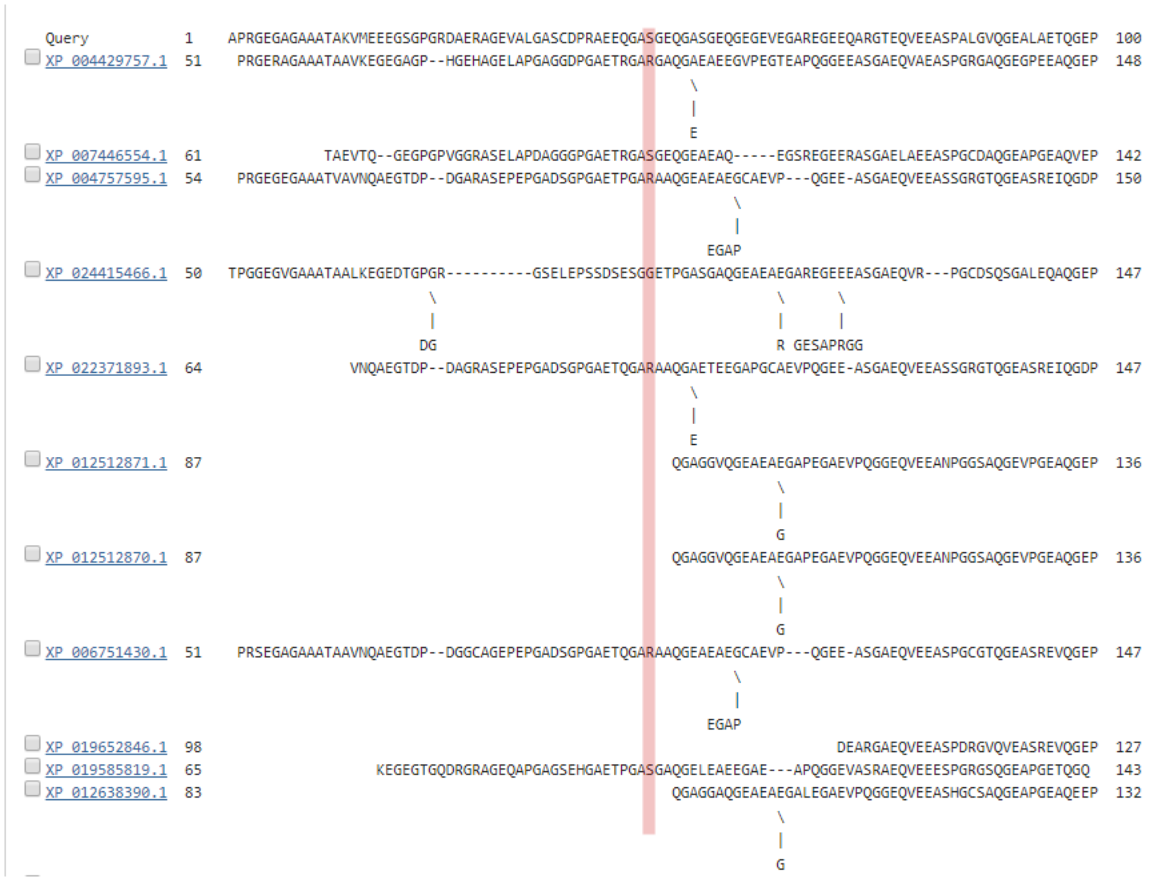
Reminder of the categories used: for category A all the infanticide pools were compared against the control pools. For category B the infanticide pool containing serial infanticide animals was compared to the controls and for capture C the pools with animals with a family history of infanticide were compared to the control pools.

In this section, the impact of SNPs located in genes was investigated. Using the SIFT algorithm [181], it was possible to highlight SNPs that could have a potential impact on the protein encoded by the gene. The SNPs classified as missense were investigated (see tables 3.37 and 3.38 for details about them). The pipeline described in section 2.11.4 was used to evaluate their potential impact. They have been divided into three sections. The first one detailed the most interesting SNPs and genes, the second genes that might be good candidate but for which there was not enough evidence to link to our phenotype of interest. Finally, the third section discussed SNPs and genes that were unlikely to be of interest. Figure 4.2 and 4.3 gave an example of the alignments used to investigate the amino acid change. Figure 4.2 showed an example of a conserved region while figure 4.3 an example of a poorly conserved region. Table 4.7 gave a summary of the SNP of interest and the target category they belong to.



```
  Query           1     FLCARLLDLRKETRAARLIQTTWRKYKLKKDLKHHQKRDKAARIIQSAVITFLTKRRLKKEVNAAVVIQKNWRRILAQRKLLMLKKEKLEKVQNQSASVI  100
□ XP_020920137.1  1256  FLCARLLDLRKETRAARLIQTTWRKYKLKKDLKHHQKRDKAARIIQSAVITFLTKRRLKKEVNAAVVIQKNWRRILAQRKLLMLKKEKLEKVQNQSASVI  1355
□ XP_023986459.1  1255  FLCARLLDLRKETRAARLIQTTWRKYKLKKDLKRHQKRDKAARIIQSAVINFLTKRRFKKEVNAALVIQKYWRRLLAKRKLLMLKKENLEKVQNKSASVI  1354
□ XP_023986460.1  1255  FLCARLLDLRKETRAARLIQTTWRKYKLKKDLKRHQKRDKAARIIQSAVINFLTKRRFKKEVNAALVIQKYWRRLLAKRKLLMLKKENLEKVQNKSASVI  1354
□ XP_007167344.1  1254  FLCARLLDLRKETRAARLIQTTWRKYKLKKDLKCHQKRDKAARIIQSAVINFLTKRRFKKEVNATLVIQKYWRRLLAKRILLMLKKEKLEKVQNQSASVI  1353
□ XP_012933393.1  1236  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKHHQERDKAARIIQSAVINFLTKQRLKKEANATLIIQKYWRRILAQRKLLMLRKEKLEKIQNKSASVI  1335
□ EHB10384.1      1243  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKHHQERDKAARIIQSAVINFLTKQRLKKEANATLIIQKYWRRILAQRKLLMLRKEKLEKIQNKSASVI  1342
□ XP_007167343.1  1254  FLCARLLDLRKETRAARLIQTTWRKYKLKKDLKCHQKRDKAARIIQSAVINFLTKRRFKKEVNATLVIQKYWRRLLAKRILLMLKKEKLEKVQNQSASVI  1353
□ XP_004685542.1  1243  FLCARLLDLRKETRAARIIQTTWRKYKLKTDLKRHQERDKAARIIQSAVITFLTKQRLKKEVNAVLTIQKYWRRLLAQRKLVMLKKEKLEKVQNKSASVI  1342
□ XP_004685541.1  1243  FLCARLLDLRKETRAARIIQTTWRKYKLKTDLKRHQERDKAARIIQSAVITFLTKQRLKKEVNAVLTIQKYWRRLLAQRKLVMLKKEKLEKVQNKSASVI  1342
□ XP_010633716.1  1278  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKHHQERDKAARIIQSAVINFLAKQRLKKEANAALIIQKCWRRALAQRKLLMLTKEKLEKIQNKSAAVI  1377
□ KFO28611.1      1136  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKHHQERDKAARIIQSAVINFLAKQRLKKEANAALIIQKCWRRALAQRKLLMLTKEKLEKIQNKSAAVI  1235
□ XP_013377331.1  1242  FLCARLLDLRKEIRAARLIQTTWRKYKLRKDLKHHQERDKAARIIQSAVINFLAKRRLIKEVNAALIIQRCWRRVLAQRKLLMLRKEKLEKLQSKSASVI  1341
□ XP_012599831.1  1231  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKRHQEQDKAARIIQSAIVNFLTKQRLKKEVNAALVIQKYWRRLLAQRKLLMLKREKLEKIQYKSASII  1330
□ XP_012599828.1  1231  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKRHQEQDKAARIIQSAIVNFLTKQRLKKEVNAALVIQKYWRRLLAQRKLLMLKREKLEKIQYKSASII  1330
□ XP_012599832.1  1231  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKRHQEQDKAARIIQSAIVNFLTKQRLKKEVNAALVIQKYWRRLLAQRKLLMLKREKLEKIQYKSASII  1330
□ XP_013013531.1  1241  FLCARLLDLRKEIRAARLIQTTWRKYKIKRDLKHLQERDKAARIIQSSVISFLAKRRLKKKVNAALIIQKCWRRVLAQRRLLMLRREKLEKIQNKSASVI  1340
□ XP_013013531.1  2649                    IQGAWRKYKAKKYLKQYLAVLKAVKIIQGYFYTKLERTRFLKMRASAVFIQRKWRAILSARHFLMMK                 2738
                                        \                                              \
                                        |                                              |
                      CKVKAACKIQAWYRCWRAR                                              IAHE
□ XP_006834297.1  1253  FLCARLLDLRKEIRAARLIQTTWRKYKMKTDLKRHEERDKAARIIQSAIINFLAKRRMRKEVNAALVIQKYWRRFFAQRKLFMLKKEKLEKIQNKAASII  1352
□ XP_012505909.1  1251  FLCARLLDLRKEIRAARLIQTTWRKYKLKTDLKRHQEQDKAARIIQSAIINFLTKQRLKKEVNAALVIQKFWRRLLAQRKLLMLNREKVEKIQYKSASII  1350
□ XP_012505908.1  1251  FLCARLLDLRKEIRAARLIQTTWRKYKLKTDLKRHQEQDKAARIIQSAIINFLTKQRLKKEVNAALVIQKFWRRLLAQRKLLMLNREKVEKIQYKSASII  1350
□ XP_023586136.1  1261  FLCARLLDLRKETRAARLIQTTWRRYKLKTDLKRYQERDKAARIIQSAVINFLTKLRLKKEIIAALVIQKYWRRFLAQRKLLMLKREKLEKTQNKSAAII  1360
□ XP_004610192.1  1265  FLCARLLDLRKETRAARLIQTAWRKYKLKTDLKRYQERDKAARIIQSAVIKFLTKQRLKRKVNAALTIQKYWRRYLAQRELLILKKEKLEKIQNKSASVI  1364
□ XP_010954610.1  1260  FLCARLLDLRKETRAARLIQTAWRKHKLREDRKRHQERDKAARIIQSAVINFLTKRRFEKEINAALLIQKHWRGLLAKRQLLMLKKEKLEKVQNKSASVI  1359
□ XP_006191149.1  1168  FLCARLLDLRKETRAARLIQTAWRKHKLREDRKRHQERDKAARIIQSAVINFLTKRRFEKEINAALLIQKHWRGLLAKRQLLMLKKEKLEKVQNKSASVI  1267
□ XP_010980509.1  1195  FLCARLLDLRKETRAARLIQTAWRKHKLREDRKRHQERDKAARIIQSAVINFLTKRRFEKEINAALLIQKHWRGLLAKRQLLMLKKEKLEKVQNKSASVI  1294
□ XP_006213310.1  1224  FLCARLLDLRKETRAARLIQTAWRKHKLREDRKRHQKRDKAARIIQSAVINFLTKQRFEKEINAALLIQKHWRGLLAKRRLLMLKKEKLEKVQNKSASII  1323
□ XP_017511590.1  1298  FLCTRLLDLRKETRAARLIQTTWRKYKLRTALKHHQERDKAASIIQSAVINFLTKQQLKKEINAALTIQKYWRRFLTKRKVLMLKKEKLEKLQHKSASVI  1397
□ XP_007987267.1  1255  FLCARLLDLRKEIRAARLIQTTWRKYKLKTDLKRHQERDKAARIIQSAVINFLAKQRLRKKVNAALIIQKYWRRVLAQRKSLILKKQKLEKVQNKAASLI  1354
□ XP_007987267.1  1347             QNKAASLIQAMWRRYRAKKYLKEYLAILKAVKIIQGCFYTKLERTWFLNVRASAIIIQRKWRAILAHEHFLMIKRHR              1446
                                        \                                           \
                                        |                                           |
                      CKVKAACKIQAWYRCWRAH                                        SAKI
□ XP_020029692.1  1254  FLCARLLDLRKEIRAARLIQTTWRKYKLRKDLQCYKERDKAARIIQSAVIKFLNKQRFKKEVNAALVIQKYWRRFLAQRKLLMLKREKLEVIQNKSASVI  1353
□ XP_007987266.1  1255  FLCARLLDLRKEIRAARLIQTTWRKYKLKTDLKRHQERDKAARIIQSAVINFLAKQRLRKKVNAALIIQKYWRRVLAQRKSLILKKQKLEKVQNKAASLI  1354
□ XP_007987265.1  1255  FLCARLLDLRKEIRAARLIQTTWRKYKLKTDLKRHQERDKAARIIQSAVINFLAKQRLRKKVNAALIIQKYWRRVLAQRKSLILKKQKLEKVQNKAASLI  1354
□ AAS55400.1      1255  FLCARLLDLRKEIRAARLIQTTWRKYKLKTDLKRHQERDKAARIIQSAVINFLAKQRLRKKVNAALIIQKYWRRVLAQRKSLILKKQKLEKVQNKAASLI  1354
□ XP_013202235.1  1233  FLCARLLDLRKEIRAARLIQTTWRKYKLKKDLKHHQERDKAARIIQSVVLNFLTRRRLRKKAHAALVIQKCWRRISAQRKLLMLKNEKLAKLQSKSASLI  1332
□ XP_021053290.1  1230  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARTIQSFVLNFLSRRRLQKKVNAALVIQKCWRRVSAQRKLLMLRNEKLAKLQSKSASLI  1329
□ XP_008829802.1  1306  FLCARLLDLRKEIRAARLIQTAWRKYKVKKDLKCHQERDKAARVIQSAAVKFLTRRRLQKKVYAALVIQKYWRRISAQRKLIMLKKEKLEKLQNKSASLI  1405
□ XP_007643382.1  1259  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARIIQSVVLSFLTRRRLQKWTSAALVIQKCWRRVLAQRQLQMLKNEKLAKLQSKSASVI  1358
□ ERE74295.1      640   FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARIIQSVVLSFLTRRRLQKWTSAALVIQKCWRRVLAQRQLQMLKNEKLAKLQSKSASVI  739
□ XP_007633990.1  1206  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARIIQSVVLSFLTRRRLQKWTSAALVIQKCWRRVLAQRQLQMLKNEKLAKLQSKSASVI  1305
□ XP_021084841.1  1233  FLCARLLDLRKEIRAARLIQTTWRKHKLKRDLKHHQERNKAARIIQSVVLSFLTRRRLQKQANAALVIQKCWRRVSAQRQLLMLKKEKLAKLQSKSASLI  1332
□ XP_006155520.1  1218  FLCARLLDLRKEIRAARLIQTMWRKYKLKTYLKYHQERDKAARIIQSAVINFLTKRRLKKEVHAVLVIQKHWRRILAQRKL----KEKLEKVQNKSASVI  1313
□ EGW10908.1      407   FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARIIQSVVLSFLTRRRLQKWTSAALVIQKCWRRVLAQRQLQMLKNEKLAKLQSKSASVI  506
□ XP_021053288.1  1230  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARTIQSFVLNFLSRRRLQKKVNAALVIQKCWRRVSAQRKLLMLRNEKLAKLQSKSASLI  1329
□ XP_021053289.1  1230  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARTIQSFVLNFLSRRRLQKKVNAALVIQKCWRRVSAQRKLLMLRNEKLAKLQSKSASLI  1329
□ AAY46816.1      1223  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARVIQSVVLNFLSRRRLQKNVSAALVIQKCWRRVSAQRKLRMLKNEKLAKLQNKSAVLI  1322
□ XP_021023769.1  1239  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKYHQERDKAARVIQSVVLNFLSRRRLQKKVSAALVIQKCWRRVSAQRKLLMLKNEKLAKLQNKSALLI  1338
□ EDL39528.1      1208  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKHHQERDKAARVIQSVVLNFLSRRRLQKNVSAALVIQKCWRRVSAQRKLRMLKNEKLAKLQNKSAVLI  1307
□ XP_021023761.1  1239  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKYHQERDKAARVIQSVVLNFLSRRRLQKKVSAALVIQKCWRRVSAQRKLLMLKNEKLAKLQNKSALLI  1338
□ XP_021023753.1  1239  FLCARLLDLRKEIRAARLIQTTWRKYKLKRDLKYHQERDKAARVIQSVVLNFLSRRRLQKKVSAALVIQKCWRRVSAQRKLLMLKNEKLAKLQNKSALLI  1338
```

Figure 4.2: Example of a conserved region, protein ASPM. Blastp results of the reference protein against the non redundant database. The substitution is deleterious, the original AA is N (asparagine, highlighted in red) and change to K (lysine). The AA and the region around it are well conserved. The top ten hits are, in order: *Sus scrofa* (Pig), *Physeter catadon* (sperm whale), *Balaenoptera acutorostrata scammoni* (minke whale), *Heterocephalus glaber* (naked mole rat) ,*Condylura cristata* (star nose mole), *Fukomys damarensis* (Damara mole-rat). Some are present twice: *Physeter catadon*: position 2 and 3*, Heterocephalus glaber:* position 5 and 6*, Condylura cristata*: position 8 and 9.



```
  Query           1     APRGEGAGAAATAKVMEEEGSGPGRDAERAGEVALGASCDPRAEEQGASGEQGASGEQGEGEVEGAREGEEQARGTEQVEEASPALGVQGEALAETQGEP  100
□ XP_004429757.1  51     PRGERAGAAATAAVKEGEGAGP--HGEHAGELAPGAGGDPGAETRGARGAQGAEAEEGVPEGTEAPQGGEEASGAEQVAEASPGRGAQGEGPEEAQGEP  148
                                                                             \
                                                                             |
                                                                             E
□ XP_007446554.1  61               TAEVTQ--GEGPGPVGGRASELAPDAGGGPGAETRGASGEQGEAEAQ-----EGSREGEERASGAELAEEASPGCDAQGEAPGEAQVEP  142
□ XP_004757595.1  54     PRGEGEGAAATVAVNQAEGTDP--DGARASEPEPGADSGPGAETPGARAAQGEAEAEGCAEVP---QGEE-ASGAEQVEEASSGRGTQGEASREIQGDP  150
                                                                                  \
                                                                                  |
                                                                               EGAP
□ XP_024415466.1  50    TPGGEGVGAAATAALKEGEDTGPGR----------GSELEPSSDSESGGETPGASGAQGEAEAEGAREGEEEASGAEQVR---PGCDSQSGALEQAQGEP  147
                                               \                                       \      \
                                               |                                       |      |
                                              DG                                       R GESAPRGG
□ XP_022371893.1  64                  VNQAEGTDP--DAGRASEPEPGADSGPGAETQGARAAQGAETEEGAPGCAEVPQGEE-ASGAEQVEEASSGRGTQGEASREIQGDP  147
                                                                             \
                                                                             |
                                                                             E
□ XP_012512871.1  87                                                       QGAGGVQGEAEAEGAPEGAEVPQGGEQVEEANPGGSAQGEVPGEAQGEP  136
                                                                                       \
                                                                                       |
                                                                                       G
□ XP_012512870.1  87                                                       QGAGGVQGEAEAEGAPEGAEVPQGGEQVEEANPGGSAQGEVPGEAQGEP  136
                                                                                       \
                                                                                       |
                                                                                       G
□ XP_006751430.1  51     PRSEGAGAAATAAVNQAEGTDP--DGGCAGEPEPGADSGPGAETQGARAAQGEAEAEGCAEVP---QGEE-ASGAEQVEEASPGCGTQGEASREVQGEP  147
                                                                                  \
                                                                                  |
                                                                               EGAP
□ XP_019652846.1  98                                                                          DEARGAEQVEEASPDRGVQVEASREVQGEP  127
□ XP_019585819.1  65                     KEGEGTGQDRGRAGEQAPGAGSEHGAETPGASGAQGELEAEEGAE---APQGGEVASRAEQVEEESPGRGSQGEAPGETQGQ  143
□ XP_012638390.1  83                                                       QGAGGAQGEAEAEGALEGAEVPQGGEQVEEASHGCSAQGEAPGEAQEEP  132
                                                                                       \
                                                                                       |
                                                                                       G
```

Figure 4.3: Example of a poorly conserved region, protein CLIC6. Blastp results of the reference protein against the non redundant database. The substitution is from serine (highlighted in red) to arginine, S to R. The AA substitution doesn't seem to be major as arginine can be seen to be present in other species. The region around the substitution is not well conserved. The top 10 hits are, in order: *Ceratotherium simum simum* (white rhino), *Lipotes Vexilifer (*Chinese river dolphin), *Mustela putorius furo* (ferret), *Desmondus rotundus* (Common vampire bat), *Enhydra lutris kenyoni* (sea otter), *Propithecus coquereli* (Coquerel's sifaka), *Leptonychotes weddelli* (weddell seal), *Ailuropoda* (panda), *Rhinolophus sinicus* (horseshoe bat). *Propithecus coquereli* is present twice (position 6 and 7).

ID	CHR	POS	Average difference	Origin	Line	Gene	Priority target
rs337246460	1	81,891,724.00	0.275(A), 0.4(B)	Cap3 cat A, B	C	TSPYL4	
rs321068974	1	81,892,394.00	0.43	Cap3 cat B	C	TSPYL4	
rs80970369	1	81,892,966.00	0.41	Cap3 cat B	C	TSPYL4	
rs327767403	1	81,908,800.00	0.39	Cap3 cat B	C	TSPYL1	
rs325683130	4	109,665,261.00	0.425	Cap3 cat B	C	CYMP	
rs339852659	4	109,671,858.00	0.455	Cap3 cat B	C	CYMP	
rs342388220	4	111,042,820.00	0.525	Cap 1 cat B	C	WDR47	**
rs332510157	4	111,098,594.00	0.5	Cap 1 cat B	C	CLCC1	*
rs321660613	4	111,116,353.00	0.274(A), 0.42(B)	Cap 1 cat A,B	C	GPSM2	**
rs342779307	5	5,742,422.00	0.292	Cap2 cat A	D	MCAT	*
rs330946522	5	5,751,299.00	0.587	Cap2 cat B	B	MCAT	*
rs340611738	6	15,036,652.00	0.4	Cap3 cat C	B	PMFBP1	*
rs333510524	6	15,036,654.00	0.4	Cap3 cat C	B	PMFBP1	*
rs333013901	9	67,604,156.00	0.2785	Cap3 cat A	C	Unknown (PCNP)	
rs339509635	9	93,078,330.00	0.455	Cap3 cat C	B	ABCB1	**
rs332900783	9	117,216,372.00	0.29, 0.29	Cap2, 3 cat A	C, B	TNN	
rs319385750	10	14,222,171.00	0.41	Cap2 cat C	B	PARP1	**
rs340769264	10	14,222,231.00	0.4	Cap2 cat C	B	PARP1	**
rs328408424	10	19,989,785.00	0.365	Cap2 cat B	H	ASPM	**
rs340828634	10	19,997,880.00	0.365	Cap2 cat B	H	ASPM	**
rs321307993	12	40,037,442.00	0.405	Cap3 cat C	B	NLE1	
rs334711032	12	50,073,425.00	0.266, 0.408	Cap3 cat A, C	B, C	ZZEF1	*
rs81438217	12	59,139,527.00	0.285	Cap 1 cat A, B	B	TRPV2	**
rs325360794	12	59,371,738.00	0.3075	Cap 1 cat A,B	C	NCOR1	**
rs81438358	12	59,376,875.00	0.28	Cap 1 cat A	C	NCOR1	**
rs692365780	12	59,418,694.00	0.3935	Cap 1 cat C	D	TTC19	*
rs346399941	13	198,251,410.00	0.47	Cap2 cat B	C	CLIC6	
rs337655527	13	203,339,714.00	0.37	Cap2 cat B	C	IGSF5	*
rs328154476	14	49,484,969.00	0.36	Cap2 cat B	C	ADORA2A	**
rs81209170	16	69,125,456.00	0.306	Cap2 cat A	C	FAM114A2	
rs81209170	16	69,125,456.00	0.306	Cap3 cat A	B	FAM114A2	
rs694857226	17	32,583,104.00	0.39	Cap2 cat B	C	AVP	**
rs324827243	17	41,286,600.00	0.46	Cap3 cat B	C	KIAA1755	*

Table 4.7: Summary table of the SNPs of interest. Average difference gives the allele frequency difference between control pools and infanticide pool(s). The origin column the capture set for which the SNP was found, line the line in which it was found. Priority shows which category it belongs to: ** is high priority target, * potential target and blank is for unlikely candidates.

4.3.5.1 Priority candidate genes and SNPs

One of the most interesting regions highlighted by the missense SNPs is located in chromosome 4 and covers four genes, *CYMP, WDR47, CLC1* and *GPSM2.* Two of these genes are good candidates, *WDR47* and *GPSM2.* Both variants were selected for line C and capture set 1 category B, and also category A for the variant within *GPSM2.*

The first SNP, rs342388220, is located in *WDR47 (*WD repeat domain 47*)* is an interesting candidate. The reference allele is an adenine (A) while the alternative is a guanine (G). In other species a G is present at this location. The allele frequencies in the infanticide animals is prevalently the alternative allele (G) while the reference control has predominately the reference allele (A). The amino acid (AA) substitution is from glutamate to lysine which is a major change from acidic (negatively charged) to basic (positively charged) in the polypeptide chain. Furthermore, the reference AA is lysine for the pig, but the blastp alignments reveal that the most prevalent AA at that location is glutamate. The region is also highly conserved. Given that the control pools have a higher frequency for the reference allele, they will have lysine at this location while the infanticide pigs will have a glutamate. The impact of this substitution is hard to predict but it could have consequences on the protein coded by this gene. While this gene and its protein are poorly characterised, the gene is highly expressed in the brain of the mouse at the embryonic stage and in adult mice. Mouse knockouts of this gene have been reported to be hyperactive. *WDR47* is also essential for brain development during the early stage of embryogenesis[220]. It is possible that a modification of the protein structure could have an impact on the mature brain.

The second candidate SNP on chromosome 4, rs321660613, is located within the *GPSM2* gene (G protein signaling modulator 2). For the pig the reference allele is a cytosine (C) and the alternative is a G. In other species the base is a G. The AA substitution is from glycine to alanine which is a change from hydrophilic to hydrophobic. Depending on where the AA is located after the folding of the protein, it could have an impact on its function. The region where the substitution is located is relatively well conserved between the species. Most animals have an alanine at this location rather than a glycine. The protein is a G-protein signalling modulator which regulates the activation of G-proteins. Other than being involved in non syndromic deafness and Chudley-McCullough syndrome, a neurologic disorder characterized by early-onset sensorineural deafness, there is not much know about it's function. However, as discussed in the previous study done on gene expression [80], other G-proteins have been linked to depression so this could be a good candidate.

Both of these candidate SNPs on chromosome 4 are in a peak region that was identified in the Quilter et al study [2], supportive that this region is linked to the genetics of maternal infanticide. These SNPs also display some of the largest changes in allele frequencies between between the control and infanticide pools, between 42 to 52%. They are all from line C which is a Large White line. These SNPs might be specific to the maternal infanticide trait in this line in particular.

The next variant, rs339509635, is located within an uncharacterised gene in the pig on chromosome 9. It has been selected for capture set 3 and category C, for line B. The reference allele is a C and the alternative is a thymine (T), most species have a C at this location. The SNP has been classified as deleterious with a change from arginine to cysteine which is a change from positively charged to neutral. The region of the protein where the AA is located is highly conserved between species. The gene at this location is unknown in the pig, but comparisons with the human and mouse return *ABCB1* (ATP binding cassette subfamily B member 1 also known as *MDR1).* The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. Its role is to transport various molecules across extra and intra cellular membranes. It is expressed in the brain in human and a GWAS has linked a SNP located in this gene to schizophrenia in a population of Ashkenazi Jews, although it did not reach genome wide significance [221]. Two others studies linked this gene to drug resistant epilepsy in Polish adult populations and in children [222, 223]. Finally a meta analysis has linked this gene to Alzheimer's Disease [224]. The fact that this gene has been associated to some brain pathologies makes this variant a good candidate.

On chromosome 10, four SNPs are classified as missense, covering two genes, *PARP1* and *ASPM.* For the two SNPs, rs319385750 and rs340769264, located in *PARP1* (poly(ADP-ribose) polymerase 1)*,* the substitution are both threonine to alanine, a change of polarity for the amino acid. The first location is in a highly conserved region of the protein. The second one is in a relatively conserved region but substitution in the protein at this location is observed in other species. The change in polarity could have an impact on the protein function, depending on how the protein structure is affected. Both SNPs were selected in capture set 2 category C and for line B. This is one of the dam lines, and category C is for the pool with a family history of infanticide. The change in allele frequency is quite high at 40%, and the infanticide pool has the alternative allele as the preferential allele in both cases. Furthermore, this region was identified in the FBAT analysis and the SNP in it reached genome wide significance. As discussed before, *PARP1* encodes an enzyme which is associated with the chromatin. It is a transferase that modifies nuclear proteins by poly(ADP-ribosyl)ation. It is involved in important cellular processes (differentiation, proliferation, recovery from DNA damage). It is also expressed in the human brain and has been linked to several brain pathologies such as Alzheimer's and Parkinson's disease [193, 192]. In the mouse it has been linked to contextual fear memory [191]. These two SNPs confirm *PARP1* as a strong candidate.

The next two SNPs on chromosome 10, rs328408424 and rs340828634, are both in the *ASPM* (abnormal spindle microtubule assembly) gene. The first SNPs, rs328408424, will probably not have a large impact on the protein, as the change it triggers does not change the polarity or charge at this location. The change is from leucine to phenylalanine. The protein region it is located in is also very variable in other species. The second SNP, rs340828634, is more interesting, it is classified as deleterious by SIFT. It triggers a change from a polar, hydrophilic and neutral amino acid (asparagine) to an hydrophobic and charged (lysine) amino acid. The protein alignment shows that the asparagine is conserved in most species. The gene *ASPM* is an ortholog of the Drosophila melanogaster "abnormal spindle" gene (*asp*). It is involved in mitotic spindle regulation which has a role in regulating neurogenesis. Microcephaly is triggered by mutations in this gene [225, 226], a GWAS study also linked it to communication disorders [227], making it a potentially interesting target. It is the only SNPs with deleterious consequence identified in line H, one of the sire lines, and both were identified in the comparison involving the serial infanticide animals. The preferential allele in the infanticide animal is the alternative allele: infanticide animals are more likely to have a charged hydrophobic lysine at this protein location.

On chromosome 12, three SNP are closely located and are within two genes, *TRPV2* (transient receptor potential cation channel subfamily V member 2) and *NCOR1* (nuclear receptor corepressor

1). The first SNP in this region, rs81438217, is in *TRPV2*. The variant has a reference allele G and an alternative allele C In most species the DNA sequences have either a A or a G when compared by phylogeny. The substitution of the AA triggers a change in the charge of the protein as it switches from histidine (positively charged) to glutamine (neutral). The region is highly conserved and most species have the neutral amino acid at this location instead of the histidine. Here the infanticide pool has a higher frequency for the reference allele, therefore a charge will be present on the protein at this location. The gene *TRPV2* encodes an ion channel, activated at high temperature. It has been linked to neural growth in developing neurons [228]. Not many functions of this gene are linked to brain pathology so far. The next two SNPs are both located in the *NCOR1* gene and have similar, but reversed, impacts on the protein AA chain. The first SNP, rs325360794, triggers a change from glycine to serine while the second, rs81438358, a change from serine to glycine. Glycine is non-polar and hydrophobic while serine is polar and hydrophilic. Interestingly, alignment of the protein sequence against other species shows that most species have the modified AA at this position (so serine and glycine respectively) in their proteins instead of the reference AA in the pig protein. Both SNPs comes from capture set 1 category A (as well as B for the first one) and were selected for line C. Capture 1 was designed based on Quilter et al [2] and *NCOR1* was one of the gene identified in this study. Unfortunately most of the known functions for this gene are linked to cancer, no studies have found any linked between this gene and brain functions so far. However it is linked to our next target, the gene *ADORA2A*, as a study [229] has found SNPs in both gene associated with gray matter volume in the cuneus component. Thus there is a potential link between this two genes.

Only one candidate SNP is located on chromosome 14, it is located in the gene *ADORA2A* (adenosine A2a receptor). The SNPs, rs328154476, triggers an AA change from glycine to glutamate which is non polar, hydrophobic to polar, hydrophilic. The region where the AA change occurs is highly conserved. The gene encodes for a member of the guanine nucleotide-binding protein (G-protein)-coupled receptor (GPCR) superfamily. It interacts with the G(s) and G(olf) family of G proteins to increase intracellular cAMP levels. It plays a role in many biological functions, including cerebral blood flow control, pain regulation and sleep. It is expressed in the brain and studies have linked it to anxiety disorder [230], mild cognitive impairment and Alzheimer's Disease [231] and neurodegeneration [232]. As mentioned previously, G proteins are of particular interest as their pathway was identified as a target in the gene expression study of Quilter et al [80] and we already have a good candidate with the SNP located in *GPSM2*. This particular SNP could also be a good candidate, it was selected in capture set 2 category B in line C.

The last candidate SNP in this priority category, rs694857226, is located on chromosome 17, within the *AVP* gene. The gene, arginine vasopressin encodes a protein of the vasopressin and oxytocin family. It is secreted in the hypothalamus and goes into the blood stream. Polymorphisms in this gene have been linked to social behaviours and psychiatric traits [233], notably depression [234] and aggression in children [235]. It also plays a role in social memory in rat [236]. Oxytocin was a protein of interest identified by the gene expression study of Quilter et al [80]. The change in amino acid, however, is valine to alanine which are both non polar and hydrophobic. The alignment of the protein show that it is relatively conserved, but most species have an alanine at this location. The allele frequency in the infanticide pool shows that only the reference allele is present, while it is less predominant in the control pools. This variant was selected in capture set 2 category B for line C. It could be an interesting candidate but it will be necessary to confirm any impact it could have on the protein function.

4.3.5.2 Potential candidate SNPs and genes

As discussed in the previous section, four SNPs of interest are present on chromosome 4, two of which are priority candidate. Another SNP in the gene *CLCC1* (chloride channel CLIC like 1), rs332510157, could be a potential candidate. It was found for line C in category B of capture set 1. In the pig the reference allele is a T and the alternate allele is a G. For other species the base at this location is a G. The AA substitution is leucine to arginine which changes the AA from neutral to positively charged. This could have an impact on the protein structure, but the alignment suggests that the structure is not well conserved for this region of the protein. Very little is known about this gene, other than it is potentially coding for a chloride channel. The only interesting fact about it is that the loss of the protein is linked to neurodegeneration [237]. As discussed in the previous section 4.3.5.1, it is in one of the peak region of the GWAS study by Quilter et al [2].

The next two SNPs are located on chromosome 5, both within the same gene: *MCAT* (malonyl-CoA-acyl carrier protein transacylase*).* They have been selected for capture set 2 for category A and B and from line D and B. The first SNP, rs342779307, might not be a very good candidate, the AA substitution is from alanine to valine and both are non polar, hydrophobic. Furthermore, the results of the blastp revealed that this location is variable, with either alanine or valine present in different species. Unfortunately the results are similar for the second SNP, rs330946522. The AA substitution is proline to leucine which are both non polar, hydrophobic. There is no good alignment when blasting the AA sequence and very little is known about the function of the *MCAT* gene. However some interesting facts are linking this gene to other candidates. The protein encoded by this gene is found exclusively in the mitochondria. It is a transacylase, which is involved in the metabolism of fatty acids. This could be an interesting target as some of or other target genes are linked to the mitochondria. The fact that two SNP are located closely together in this gene is important, this SNP and other could define haplotypes.

The next two SNPs, rs340611738 and rs333510524, are located on chromosome 6 and both are within the *PMFBP1* (polyamine modulated factor 1 binding protein 1) gene, just two nucleotides apart. The SNPs have an A and a T as reference and both have a G as their alternative allele. Both SNPs affect the same amino acid, the first one trigger a change from serine to arginine which changes the charge from neutral to positive. The second is a change to glycine, from polar to non polar, so hydrophilic to hydrophobic. If both SNPs have the alternative allele, the substitution is also serine to arginine. The region where the nucleotides and amino acids are located is highly variable. The gene itself is expressed in the brain in human at a low level but nothing is known about the function of the protein it produces. As this region came up in the section looking at region with a high number of variants above our threshold 4.3.6, it is a region that should be investigated in more detail.

The SNP on chromosome 12, rs334711032, is within the *ZZEF1* (zinc finger ZZ-type and EF-hand domain containing 1) gene. The same SNP was selected on capture set 3, category A for line B, and category C for line C. Both lines are the dam lines and capture set 3 is based on the PO study. This might be a good candidate for dam to daughter preferential transmission. The allele frequency difference between control and infanticide for line C is quite high at 40%. The SNP alleles trigger a change from alanine to serine which is a change from non polar, hydrophobic to polar, hydrophilic. The region where the substitution is located in the protein is highly conserved: the alanine is present in most species. Not a lot is known about the gene function but it is highly expressed in the hippocampus and has been linked to impaired memory in the mouse [238], making it a candidate.

Another potential candidate SNP, rs692365780, is located on chromosome 12, within the gene *TTC19* (tetratricopeptide repeat domain 19). This gene is an interesting target because its function has been linked to ataxia [239], a neurological disease characterised by a lack of coordination in muscle movements. More importantly it is related to mitochondrial complex III deficiency and neurological impairment [240]. As we saw in the previous studies (section 1.4.2), mitochondrial dysfunction has been associated with bipolar disorder and schizophrenia [66, 67]. While the gene is a good candidate, unfortunately the SNP is not. The substitution it triggers switch an isoleucine for a valine, which are both neutral and non polar. Furthermore the protein region is not very conserved . Despite of this, it might be interesting to investigate the gene in more details, as other variants might be more impactful.

On chromosome 13, one of the potential candidate, rs337655527, is located within the gene *IGSF5* (immunoglobulin superfamily, member 5). The SNP has been selected for capture set 2 and category B (serial infanticide) on line C. This gene has been linked by GWAS with suicide attempts in mood disorder patients [241]. The protein region is highly conserved but the change is valine to alanine: both are polar, hydrophobic. The alignment returns a prevalent alanine in that location so the valine could have an impact even if the polarity is similar. However this is difficult to assess.

The last SNP in our list of potential candidates, rs324827243, is located within the *KIAA1755* gene. Not a lot is know about the gene but it is expressed in the brain. The change in AA is from alanine to threonine which is change of polarity, from polar, hydrophilic to non polar, hydrophobic. The AA is located in a highly conserved region and alanine is present at this location in all of the protein alignments. It could be a good candidate due to difference in allele frequencies. It is found in capture set 3, category B and in line C. The aggressive pool has only the reference nucleotide C while the control have a allele frequency for the reference allele of 56% for both. It is one of the largest change in allele frequencies in the table. The function of the gene needs to be investigated to validate this candidate.

4.3.5.3 Unlikely candidates.

The first group of SNPs, located on chromosome 1, are within two different genes, *TSPYL4* (testis-specific Y-encoded-like protein 4) and *TSPYL1* (testis-specific Y-encoded-like protein 1). There are three SNPs with missense consequence in *TSPYL4,* rs337246460, rs321068974 and rs80970369. One in *TSPYL1,* rs327767403. There several changes in the amino acid chain of the protein coded by the gene *TYSPL4.* The three SNPs are causing a change of polarity at that location, with changes from isoleucine to threonine (non polar to polar), threonin to alanine (polar to non polar) and finally from glutamine to proline (polar to non polar). The blastp of the protein sequences around the AA subsitution shows that the protein is not very conserved between species at these locations. As the region is not very conserved it is hard to predict if a change of polarity will affect the protein function. Furthermore there is very little known about this gene. The change of polarity also happens for TSPYL1*,* a change from proline to serine. For this protein the region is conserved and most species have conserved the proline. This substitution could have an impact on the protein function but very little is known about it. The *TSPYL* family is related to genes located on the Y chromosome and involved in male fertility. However one member is known to linked to brain function *TSPYL2.* These variants were found in line C, one of the dam lines, and were picked up by the capture set 3, category B (serial infanticide), suggesting a potential dam to daughter transmission. In all cases the alternative allele was more frequent in the infanticide pool than in the control pool. Therefore the substitution of polarity in the protein happens more often in infanticide animals.

Of the candidates found on chromosome 4, two SNPs are less likely to be good candidate due to the unknown nature of the gene they are in, they are rs325683130 and rs339852659. They were found for capture set 3 category C (family history of infanticide) and one of them is classified as deleterious by SIFT. The gene at this location (109.6MB) is unknown in the pig but comparisons with the human returns *CYMP* (chymosin pseudogene) as a potential match for this gene*.* The pig gene is classified as a pseudo gene and has to potential to produce a protein. The protein might have divergent functions to *CYMP.* The first AA substitution is a like for like substitution as it is change from serine to asparagine, both are neutral and polar. The second is deemed deleterious by the SIFT prediction however the change is between two neutral and non polar AAs, leucine to proline. The blastp alignments gave little information as there were no matches for the part of the protein where the AA substitution happens. Unfortunately, there is very little known about the gene other than it is a peptidase, it is not well characterised and the only indication of function is a GWAS study on depression [242] that identified one SNP, located in this gene. It almost reached genome wide significance (5.10^{-7}).

The next variants in this category are located on chromosome 9. They cover an uncharacterised gene and *TNN* (tenascin N*).* The first SNP, rs333013901, is classified as deleterious by the SIFT prediction. The difference in allele frequencies is modest, the infanticide pools being at 50% and 41.7% and the controls at 22% and 14%, the average difference is 28% between the two categories of pools. This SNP was also only identified for capture set 3, category A in line C. It is probably below the filtering threshold for category B and C. However the change in the AA chain is interesting, it is a change from threonine to arginine which modifies the charge of the amino acid, from neutral to positively charged. The position is also well conserved between species. However the gene in the pig is very short and not characterised. Comparing it to the human and mouse genome suggest that it might be *PCNP* (PEST proteolytic signal containing nuclear protein). Unfortunately the gene *PCNP* is poorly characterised and not much is known about its function other than being involved in the cell cycle [243].

The other variant on chromosome 9, rs332900783, is in the gene *TNN* (tenascin N). It has an A as its reference allele and a G as its alternative allele. Other close species (cow, horse) have also a A at this position but other have a G. The change in amino acids is isoleucine to valine which are both neutrally charged and non polar. *TNN* itself is uncharacterised and has low expression in the mouse and human brain, however tenascin proteins are involved in axon generation and the protein encoded by this gene has been found in the hippocampus [244].

The SNP located in the gene *NLE1* (notchless homolog 1) on chromosome 12, rs321307993, might not be an interesting candidate. Nothing is known about this gene's function but it is expressed in the brain. The change in AA triggered by the SNPs is likely to have little impact as the substitution has the same polarity, polar and the same charge, positive. The position is however highly conserved unfortunately, as little is known about this gene and its protein in other species, it makes it hard to evaluate this SNPs.

One SNP located on chromosome 13, rs346399941, is located within the *CLIC6* (chloride intracellular channel 6) gene. The AA substitution triggers a change in charge in the protein, from serine to arginine, from neutral to positively charged. Unfortunately the AA chain is not very conserved at this location and nothing relevant to a role in the brain is known about this gene.

Another variant in this category, rs81209170, is located on chromosome 16 and within the *FAM114A2* (family with sequence similarity 114 member A2) gene. Unfortunately despite begin found in two different capture sets (two and three, both for category A), it is not a high priority candidate. The change is from lysine to arginine, both are hydrophilic and positively charged. The region is relatively conserved, but both lysine and arginine are presents at that location in other species. Furthermore little is known about the function of this gene product.

Most of our candidate SNPs were found in line B and C, the two dam lines. This suggests that the genetic component of maternal infanticide is more prevalent in the line B and C, which is interesting as these two lines are the pure breed lines. One hypothesis could be that pure breed lines are more susceptible to the influence of the genome compared to cross breed lines.

4.3.6 Regions of interest

Reminder of the categories used: for category A all the infanticide pools were compared against the controls. For category B the infanticide pool containing serial infanticide animals were compared to the controls and for capture C the pools with animals with a family history of infanticide were compared to the controls.

In order to identify more regions of interest, the repartition of the SNPs passing the filter threshold were investigated using genomic plots. By visually inspecting the genomic plots generated for variants passing the threshold and overlapping with target regions, it became clear that some regions are covered by a large number of variants.Other regions only have a few variants covering them. Regions with a large number of SNPs passing the threshold or with genes of interest were selected in order to investigate the genes present in that region.

The category investigated are category B and C as A tend to select the regions in common between the two, usually with a lower difference in allele frequencies, resulting in lower amount of SNPs selected. One of the aim of this work is to try to find genes or region that are more likely linked to animals classified as serial infanticide or relating more to animals with a family history of infanticide. Therefore the region was first investigated for category B and C before looking at its presence or not in category A.

4.3.6.1 Capture set 1

The first region of interest is on chromosome 3 at 27MB, it is covering the gene *XYLT1* (xylosyltransferase 1) for category A and B mainly and to a lesser extend on category C. In category B, 70 variants cover this region, 52 are found in category A, but only 2 for category C. This region was one of the target region in the GWAS study by Quilter et al [2]. The SNP observed were from line C mainly. Unfortunately there are no known functions linking *XYLT1* to our phenotype of interest so far, but it is expressed in the brain.

The second region of interest is more interesting and is located on chromosome 3. It covers the gene *RBFOX1* (RNA binding fox-1 homolog 1) and its upstream region. It is found for all three category, but mainly for line C again. For category A, SNPs in lines B and D are also found. The number of variant passing threshold in this region is large, 321 for category B, and 310 for category C and only 70 for category A. This region was one of the major peaks identified in the Quilter et al study [2]. It is of particular interest because it is syntenic with an human locus on chromosome 16 that has been linked to puerperal psychosis [108] and bipolar disorders [109]. The gene is also almost exclusively and highly expressed in the human brain. The region surrounding it has been recently linked to depression by GWAS [245, 246]. It has also been implicated in other psychiatric and brain disorders such as attention-deficit hyperactivity disorder (ADHD) [113, 115, 114], autism [247, 112], generalised anxiety disorder [248], and copy number variations in this region are linked to developmental coordination disorder [249]. The protein produced by the gene is one of a Fox-1 family of RNA-binding proteins that is evolutionary conserved. It is also linked to *GRIN1* (glutamate ionotropic receptor NMDA type subunit 1*),* a target of interest from previous studies. *RBFOX1* modulates the expression of alternate splice variants for neuronally expressed genes, including *GRIN1 [250].* This region is a very strong candidate due to the fact that it is found in all three categories. It is most strongly correlated to line C (Large White), but the fact that some other lines were found in different categories is encouraging. As no missense SNPs were found in this gene, the variants are probably located in introns or in the 5' region of the genes. This suggest that any influence they have on the gene would come from splice variants or regulatory features of the genes, thus regulating its expression rather than changing its protein. A more detailed study of this region could be lead to some interesting results.

One of the region with a very large number of SNPs covers three of the candidate genes identified in the previous section, on chromosome 4 at 111MB, *WDR47, CLCC1* and *GPSM2.* Once more most of these SNPs are selected for line C, mainly category A and B, with only a small fraction found for category C. For category B 309 variants are selected, 114 for category A but only 6 for category C. Line C is the large white pure breed. As the SNPs are found in majority for the category B, it suggests an influence on the fact that the animals will be those with serial infanticide trait. As discussed before, all three genes in the interval are interesting candidates.

The next region is on chromosome 12 and is another region that confirms some of our previous results. It covers three of the genes identified as potential candidates before, *NCOR1, TRPV2* and *TTC19.* This time, some SNPs above the threshold are found in line B, but in the majority are in line C. The region is well covered with SNPs passing the threshold for categories A and B but has less coverage for category C. The number of variants passing the threshold for category A is 249, 124 were selected for category B and only 23 for category C. It does confirm this region as a candidate region for further investigation.

A region on chromosome 14 with a large number of SNPs above the thresholds. For category A 248 selected variants are present from line B and C. For category B only 70 are present, all from line C. Finally, for category C 154 variants are selected from both line B and C. The intervals covers four genes, *POLE, GALNT9, P2RX2* and *FBRSL1.* Of these four, one of them is a more interesting candidate, *P2RX2.* This gene is linked to the regulation of the release of vasopressin and oxytocin [251] . One of the gene of interest selected by the missense SNPs analysis in the previous section was *AVP,* which encodes for polypeptides of the vasopressin and oxytocin family. Low levels of oxytocin have been linked to aggressive behaviour [83] and, as described in the previous section, other psychiatric disorders are linked to both vasopressin and oxytocin. The later hormone was also a protein of interest in Quilter et al gene expression study [80], making this key target. The other genes in this region have no function linked specifically to the brain or behavioural phenotypes but they are all expressed in the brain. This region was also one of the top haplotype region found in Quilter et al [2].

The last region of interest for capture set 1 is located on chromosome 15 and cover the gene *PAX3* (paired box 3). This region was selected for all three categories, but once more only found in line C. The number of variants passing the threshold for category A is 99, 43 for category B and 49 for category C. The 3' region of this gene and the gene itself are one of the regions that reached genome wide significance in the GWAS carried out by Quilter et al [2]. The gene is critical in foetal development and more specifically neurological development. In mice mutations of this gene cause spina bifida and exencephaly [123]. In humans, mutations in this gene are linked to major developmental syndromes, such as the Waardenburg syndrome [252] and the craniofacial-deafness-hand syndrome [124].

4.3.6.2 Capture set 2

For category B (serial infanticide) of capture set 2, several regions confirm some of the results found using the missense SNPs table. A large number of SNPs are found to cover the region of the *MCAT* gene, on chromosome 5 at 5.5MB, for line B. For category A only 8 variant were selected, 71 were present for category B and 3 for category C.

Another region covered by a high proportion of SNPs, from line H, covers the *ASPM* gene region, on chromosome 10 at 20MB. Only 2 and 4variants were selected for category A and C respectively in line B and D. For category B, 47 variants passed threshold for line H. Another interval is on chromosome 14 and covers the *ADORA2A* gene. Here the SNPs are all selected for line C. No variants were selected for category A and B, 9 are selected for category B but the target interval is really small. Despite the low number of variant this region should be investigated further. For category C animals (family history of infanticide), a large number of SNPs (47) cover the gene *PARP1,* which was also one of our targets from the missense SNPs, this time all the SNPs were selected from line B.

Other regions not found in the previous analysis but covered by a large number of SNPs are described below. On chromosome 3 there is a region covering the gene *TGFA* (transforming growth factor alpha*),* selected for category B and in line C, with 113 variants selected. Unfortunately there are no know function for this gene linking the gene to our phenotype of interest to date. The next region of interest found for category B with 132 selected variants covering the gene *ARPP21* (cAMP regulated phosphoprotein 21), on chromosome 13 at 21 MB. Most of the SNPs selected are found in line C. The same region is also selected for category C with a larger proportion of SNPs above threshold, almost all found in line B, but two in line C. This gene is a very promising candidate as it encodes for cAMP-regulated phosphoprotein. It is highly expressed in the brain and enriched in the caudate nucleus. It is an area of the brain involved in memory, sleep and the reward system, which is linked to dopomine secretion. In the mouse a similar protein regulates the effects of dopamine in the basal ganglia. As discussed before, dopamine plays a important role in the prevention of depression and is a target for several of the genes identified in this study and previous ones. It is also a regulator of calmodulin signalling and when a mouse knockout was generated, it resulted in anxiety like behaviours [253]. Deletion of this gene could be a potential pathogenic factor for intellectual disability [254]. This makes this region a priority candidate in line C.

The final region of interest for capture set 2 is located on chromosome 11 and covers the gene *LRCH1* (leucine rich repeats and calponin homology domain containing 1), it was selected on line B for category C with 626 variants present. This gene was found in a RNA sequencing experiment performed in our lab and part of another thesis [176]. The gene is expressed in the brain but so far no interesting function related to our phenotype has been found. However it is classified as a negative regulator of GTPase activity, which links it to G-proteins.

4.3.6.3 Capture set 3

The first region of interest for capture set 3 is found on chromosome 1 for categories A and B (serial infanticide) in line C, and covers the genes *TYSPL4* and *TYSPL1.* As discussed previously (section 4.3.5.3) this region had several missense SNPs in it and could be a candidate for further investigation, however the genes function is currently not well characterised. The number of selected variants for this region for category A is 50, for category B it is 78 and for category C, there are only 3 variants.

The next region of interest is located on chromosome 2 at 84.5MB and covers two genes, *COL4A3BP* (collagen type IV alpha 3 binding protein) and *POLK* (DNA polymerase kappa). It was selected for category C (history of infanticide) and found in line B, 116 variants passing threshold are present. Both genes are expressed in the brain and mutations in *COL4A3BP* have been linked to developmental disorders in a large scale study [255]. It is also involved in the Goodpasture syndrome, an auto-immune disorder. *POLK* functions are largely linked to cancer. This region could be interesting but there is no clear link to our phenotype of interest.

Another region with a tight cluster of SNPs is located on chromosome 3 and was selected for category C, on line B. A total of 180 selected variants are found in this region. Two genes are close to this location, *RNF144A* (ring finger protein 144A) and *RSAD2* (radical S-adenosyl methionine domain containing 2). Both are expressed in the brain but no functions are described that would overlap our phenotype of interest.

The next region is on chromosome 4 and was selected for the three categories, A, B and C. Interestingly both categories were selected on different lines. For category A , 115 SNPs were found from line C. For category B, 100 SNPs were selected and came for line C. For category C, 65 SNPs were and they were selected for line B. The gene this region covers is *ASAP1* (ArfGAP with SH3 domain, ankyrin repeat and PH domain 1). It is expressed in the brain and a human GWAS study has linked it schizophrenia in the Ashkenazi Jews [221]. This could be interesting as the Ashkenazi Jews are a population which had very little genetic mixing. Commercial breeds tend to also have poorer genetic mixing, due to the selection for particular trait. This lack of genetic mixing will elevate the frequencies of certain alleles in the population, some of which could have deleterious effects.

There are three other regions of interest on chromosome 4. The first two were selected for category B for line C, the first one with 242 variants and covering *NKAIN3* and the second with 116 selected variants and covering *KCNA2, KCNA10* and *PROK1.* The second region was also found for category A, with 66 SNPs selected originating from line B and D. The last region was selected for category C and line B, covering *EXT1* with 124 variants. The first candidate *NKAIN3* (sodium/potassium transporting ATPase interacting 3) is highly expressed in the brain but little is known about its function apart from one GWAS on Alzheimer's disease which found a genome wide significant SNPs in this gene [256]. For the second region, found in categories A and B, only *KCNA2* (potassium voltage-gated channel subfamily A member 2) might be an interesting candidate. It is expressed in the brain and mutations in this gene have been linked to episodic ataxia and hereditary ataxia [257, 258]. Ataxia is a neurological disorder that affects coordination, balance and speech in humans. The other genes in this region (*KCNA10* and *PROK1*) have no functions linked to the brain or any psychiatric disorders. The final gene of interest on chromosome 4, *EXT1* (exostosin glycosyltransferase 1) is expressed in the brain. So far, most of its functions are linked to bone growth and related diseases. The allele frequency differences in this region of *EXT1* are very high. If this gene was to be linked to any psychological disorder or brain functions, it could be a good candidate.

One of the gene identified in the missense SNPs table and confirmed in the region study is *PMFBP1.* It is located on chromosome 6 and is covered by a cluster of SNPs with a relatively large allele difference between control and infanticide pools (40 to 60%). A total of 100 SNPs were found in line B and were selected for category C (family history of infanticide). A second region of interest on chromosome 6 is just downstream of the gene *CDH11* (cadherin 11). The SNPs, at total of a 172, are also from line B, and were selected for category C. One GWAS found genome wide significant association of this gene with schizophrenia [210].

The next genomic interval is located on chromosome 9. It covers the region upstream and the gene *EZH2* (enhancer of zeste 2 polycomb repressive complex 2 subunit) itself. This region was selected for categories A (35 SNPs), B (serial infanticide, 32 SNPs) and C (family history of infanticide, 213 SNPs), but on different lines: on line H for category B, line B for category C and line B and C for category A. This gene has been linked to ataxia [259] and schizophrenia [260]. It is the only region which has been selected for line H, one of the two lines with the highest incidence of maternal infanticide.

Two regions have been selected on chromosome 10. One is upstream and covering partially the gene *FRMD4A* (FERM domain containing 4A), and the second one is within the *GPR158* (G protein-coupled receptor 158) gene. For the gene *FRMD4A*, the region has a large cluster of SNPs for category A (128 SNPs), B (44 SNPs) and C (255 SNPs) but for different lines, in line C for category B and line B for categories A and C. The gene itself has been associated to Alzheimer's disease by GWAS, and the variant associated reached genome wide significance [261]. The second gene, *GPR158*, also has a large cluster of SNPs for both category B (84 SNPs) and C (353 SNPs), but again for different lines, C and B respectively. This gene is almost exclusively expressed in the brain and the protein it encodes is a G protein-coupled receptor. One study found that it regulates stress induced depression in the mouse and increased expression of this gene has been observed in the pre-frontal cortex in depressive animals [262]. Once again, it is a gene linked to G-protein signalling, linking it with several of our other target regions.

One interesting observation from this set of data is that line B and C are still the predominant lines for which regions surpass our threshold. More specifically, the lines are segregating between category B (line C) and category C (line B). This suggests that line B, a Landrace line, might have a more strong genetic linkage for the pathology in category C, which is for the pools with a family history of infanticide. For line C, a large white line, the predominant category is B, linked to the serial aggressor individuals. It suggests that the genetic component for this line might be different, and the breeds as a more systematic occurrence of infanticide in individuals. Disappointingly there are very few results for either line D or H. This is true for all capture sets. Different genes might be involved for these lines, or the regions we selected to study are not the best candidate regions for these two lines.

5 Conclusion and further work

This work provides a good example of how the genetic tools currently available can be used to investigate the genetic components of a pathology. The staged screening process used in this work could be applied to other diseases and other species. Using a combined approach of genotyping and sequencing opens new possibilities to identify the genes and loci in the genome linked to maternal infanticide. The results presented here suggest that the genetic component of this pathology is not restricted to a single location of the genome, but that several loci are contributing to an increased risk of infanticide. Each of these regions will contribute modestly to the overall phenotype, but taken together, they can increase the risk of incidence of infanticide events. Regions identified might be shared by different lines, others are more specific to a single line, suggesting variability of the genetic component in different lines and breed. It is possible that pressure of selection due to commercial breeding has the unwanted consequence of selecting different susceptibility loci in different lines and breeds. This could be caused by the LD structure present in the genome. The desirable traits selected for breeding might be linked by LD to less desirable trait. Thus these traits will "hitch-hike" with the selected trait and become more prevalent in a population selected for breeding. It was also noted that some lines had more candidates in one of our three categories. Line C, the Large White, has more candidates found in the category including the serial infanticide sows. Line B, the Landrace, on the other hand, has more candidates for the category involving the animals selected because of their family history of infanticide. Line D and H returned very few regions of interest. It is difficult to understand why, one hypothesis could be that because they are cross breeds, their genomes are more heterogeneous and therefore the penetrance of causal loci is lower than the pure bred lines. There are some common trends between some of the genes that were identified as targets by previous studies and this work. Genes involved with G-protein pathway, the regulation of dopamine, vasopresine and oxytocin levels were found in different parts of this work. Another interesting trend the are genes involved with the mitochondria and energy production, suggesting that alteration of the genome impacting on the way the cells produce energy could have an behavioural impact.

While looking at individual gene function can lead to identifying interesting targets, genes tend to work together in pathways and networks to regulate a biological functions or processes. Therefore, the list of target genes obtained can be investigated as group of genes using tools such as DAVID [179]. It uses an approach called gene set enrichment, comparing a list of genes of interest against the background of all the genes in the genome to identify functions that may be over-represented in the candidate genes compared to all the genes. This analysis was performed using human genes as they are better annotated. Unfortunately no significant enriched clusters were found, only a few genes grouped with related functions, such as *ADORA2A* and *AVP* involved in glutamate secretion and *NAV2, NRN1, PAX3* and *RBFOX1* which are involved in the development of the nervous system. Both type of functions are linked to the nervous system which is encouraging. However the lack of any other interesting function linking our candidate genes together is disappointing. This could be due to poor characterisation of a lot of our candidate genes, more work is needed to expand the knowledge of the functions linked to these genes.

In order to confirm some of this work and the region and genes identified, restriction fragment length polymorphism (RFLP) work has started. This will confirm if the difference in allele frequencies found a certain loci in the pools is valid for single individuals. Using this approach, genotype on individual animals can be determined and used to determine the type of model they contribute to the trait, such additive, dominant, etc... It can be done for a lot of animals at a low cost. This work will be carried out on selected loci for which primers can be designed. Another possibility to validate some of our variants is to investigate which allele is present in lines from the same breed where no or lower incidence of maternal infanticide was reported.

If the RFLP approach validates some of the candidates found here, more experimental work will be needed. Most of the variants selected are in non coding regions, to investigate their impact on the function of genes, cell models or animal models will need to be designed. Without validation, the findings of this thesis will remain theoretical.

The approach used in this study returns some interesting results but is not without flaws. For example the array used has low coverage compared to other arrays available now. Furthermore, not all the SNPs covered are of interest for all breeds. While the design used five different breed to select SNPs, the state of the reference was poor at the time of its conception and a lot of SNPs were of poor quality. Given the improvement both in array technology and in the reference genome, it would be possible to design a much better array. These advances can help increase the number of probes on the array, to similar density used in human for example. Furthermore, with more sequencing data becoming available every day, probes specific to breeds could be integrated to broaden the range of lines and breeds that can be investigated using it. For this study, the array design used was based on version 10.2 of the genome, redesigning it for version 11 of the genome might refine some of the regions. This could lead to better typing of all the LD blocks present in the pig and get a better understanding of their impact on this phenotype. Some studies have used this approach to look at heritability of traits in families, by typing LD blocks using arrays in human [263].

This is also true for the sequence capture design; some of the probes are also targeting small areas and this could result in missing some interesting target loci. Some genes might have been left out of the capture probe design because of the quality of the genome build 10.2. Now that a better genome build is available, it would be worth checking if more genes could be included. Another approach to resolve this issue is to design PCR primer targetting the genes of interest and produce amplicons for sequencing. This is a low cost solution, allowing the sequencing of the target regions in a large number of animals.

Pooling animals was a sensible approach because of the cost related to sequencing at the time, but unfortunately most of the infanticide pools had a low number of animals in them, raising the question of how representative they might be as a sample of the infanticide population. Obtaining good samples is also difficult, especially for some of the categories, such a serial infanticide. Breeders tend to kill sows for economic reasons, as losing a litter is a significant financial loss. If good samples can be provided, the approach described here could be revisited using single animals instead of pools. This is provided that sequencing becomes affordable enough, and a sufficient number of samples can be collected. This would allow the use of more rigorous statistical methods to compare the infanticide animals against the controls. Another approach that could follow from this work is to sequence all the members of several families in order to conduct a linkage study based on variants called via sequencing. Using single individuals instead of pools would also enable the calling of structural variants (or copy number variant), which was not possible due to the algorithm used to call variants on the pools. These approaches could be done on the whole genome if enough funding is available, or using sequence capture to lower the costs. An hybrid approach could also be to sequence a few animals at genome wide level to identify new regions of interest before designing a capture panel to interrogate these in more detail.

If a new array can be constructed, or if enough sequencing can be financed, combining family and population analysis can be a powerful tool that has been used with success [264]. Furthermore, combining a new array and a sequencing approach could prove to be a powerfull, especially for difficult diseases [265].

Finally, another avenue of investigation would be to look at the epigenetic change between control and infanticide animals, for example comparing methylation levels of the DNA between the two groups. Ideally the methylation change should be investigated in the brain, it might help to shade light on some of our candidates for which no functions related to the brain or its pathologies were found. This could be combined with a gene expression study to assess the impact of the methylation of the genome on the expression of the genes.

While no definitive targets were found using our approach, several interesting regions have been identified for further work. With the advances in technology and refinement of the pig genome, it will be soon possible to confirm and potentially find additional targets with greater certainty and at a lower cost. The approach used here is not restricted to the pig and maternal infanticide, it could be applied to other pathologies and animals.

Contribution to work related to this thesis

An association and haplotype analysis of porcine maternal infanticide: A model for human puerperal psychosis?

Quilter, C. , Sargent, C. , Bauer, J. , Bagga, M. R., Reiter, C. P., Hutchinson, E. L., Southwood, O. I., Evans, G. , Mileham, A. , Griffin, D. and Affara, N. (2012),

Am. J. Med. Genet., 159B: 908-927. doi:10.1002/ajmg.b.32097

Bibliography

[1] D.C. Lay Jr., R. L. Matteri, J.A. Carroll, T.J. Fangman, and T. J. Safranski. Preweaning survival in swine. *Journal of Animal Science*, 80:E74–E86, 2015.

[2] C.R. Quilter, C.A. Sargent, J. Bauer, M. R. Bagga, C. P. Reiter, E. L. Hutchinson, O. I. Southwood, G. Evans, A. Mileham, D.K. Griffin, and N.A. Affara. An association and haplotype analysis of porcine maternal infanticide: A model for human puerperal psychosis? *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159B(8):908–927, 2012.

[3] H.A.M. Van der Steen, L. R. Schaeffer, H. de Jong, and P. N. de Groot. Aggressive behavior of sows at parturition. *Journal of Animal Science*, 66(2):271–279, 1988.

[4] P.W Knap and J.W.M Merks. A note on the genetics of aggressiveness of primiparous sows towards their piglets. *Livestock Production Science*, 17(0):161 – 167, 1987.

[5] Congying Chen, Colin L. Gilbert, Guangcheng Yang, Yuanmei Guo, Anne Segonds-Pichon, Junwu Ma, Gary Evans, Bertram Brenig, Carole Sargent, Nabeel Affara, and Lusheng Huang. Maternal infanticide in sows: Incidence and behavioural comparisons between savaging and non-savaging sows at parturition. *Applied Animal Behaviour Science*, 109:238 – 248, 2007.

[6] Gonyou H.W. Harris M.J. Savaging behaviour in domestic gilts: A study of seven commercial farms. *Canada Journal of Animal Science*, 85:435–44, 2003.

[7] Gonyou H.W. Harris M.J., Bergeron R. Parturient behaviour and offspring-directed aggression in farmed wild boar of three genetic lines. *Applied Animal Behaviour Science*, 74:153–163, 2001.

[8] Forde J.N.M. Piglet- and stockperson-directed sow aggression after farrowing and the relationship with a pre-farrowing, human approach test. *Applied Animal Behaviour Science*, 75:115–132, 2002.

[9] K. A. McLean, A. B. Lawrence, J. C. Petherick, L. Deans, J. Chirnside, A. Vaughan, B. L. Nielsen, and R. Webb. Investigation of the relationship between farrowing environment, sex steroid concentrations and maternal aggression in gilts. *Anim. Reprod. Sci.*, 50(1-2):95–109, Feb 1998.

[10] Congying Chen, Colin L. Gilbert, Guangcheng Yang, Yuanmei Guo, Anne Segonds-Pichon, Junwu Ma, Gary Evans, Bertram Brenig, Carole Sargent, Nabeel Affara, and Lusheng Huang.

Maternal infanticide in sows: Incidence and behavioural comparisons between savaging and non-savaging sows at parturition. *Applied Animal Behaviour Science*, 109:238–248, 2008.

[11] I.A. Sneddon V.E. Beattie, N. Walker. Influence of maternal experience on pig behaviour. *Applied Animal Behaviour Science*, 46:159–166, 1996.

[12] I.A. Sneddon V.E. Beattie, N. Walker a. An investigation of the effect of environmental enrichment and space allowance on the behaviour and production of growing pigs. *Applied Animal Behaviour Science*, 48:151–158, 1996.

[13] Richard B. D'Eath and Helena E. Pickup. Behaviour of young growing pigs in a resident-intruder test designed to measure aggressiveness. *Aggressive Behavior*, 28(5):401–415, 2002.

[14] Congying Chen, Guo, Yuanmei, Yang Guangcheng, Yang Zhuqing, Zhang Zhiyan, Yang Bin, Yan Xueming, Perez-Enciso Miguel, Ma Junwu, Duan Yanyu, Brenig Bertram, and Huang Lusheng. A genome wide detection of quantitative trait loci on pig maternal infanticide behavior in a large scale white duroc erhualian resource population. *Behav Genet*, 39(2):213–219–, 2009.

[15] I.A. Sneddon V.E. Beattie, N. Walker. Effect of rearing environment and change of environment on the behaviour of gilts. *Applied Animal Behaviour Science*, 46:57–65, 1995.

[16] Peter Lovendahla, Lars Holm Damgaarda, Birte Lindstrbm Nielsena, and Lotta Rydhmer Karen Thodberga, Guosheng Sua. Aggressive behaviour of sows at mixing and maternal behaviour are heritable and genetically correlated traits. *Livestock Production Science*, 2005.

[17] Claire R. Quilter, Sarah C. Blott, Anna E. Wilson, Meenakshi R. Bagga, Carole A. Sargent, Gina L. Oliver, Olwen I. Southwood, Colin L. Gilbert, Alan Mileham, and Nabeel A. Affara. Porcine maternal infanticide as a model for puerperal psychosis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 144B(7):862–868, 2007.

[18] T. H. Morgan. Random segregation versus coupling in mendelian inheritance. *Science*, 34(873):384, Sep 1911.

[19] R. C. Punnett. Linkage in the sweet pea (lathyrus odoratus). *Journal of Genetics*, 13:101–123, 1923.

[20] Ken-ichi Kojima R. C. Lewontin. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960.

[21] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, 32(3):314–331, May 1980.

[22] E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 84(8):2363–2367, Apr 1987.

[23] Catherine M Hearne, Soumitra Ghosh, and John A Todd. Microsatellites for linkage analysis of genetic traits. *Trends in Genetics*, 8(8):288–294, 1992.

[24] N.E Morton. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 1955.

[25] R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Hum. Hered.*, 21(6):523–542, 1971.

[26] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, 58(6):1347–1363, Jun 1996.

[27] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nat Protoc*, 6(2):121–133, Feb 2011.

[28] M. Durner and D. A. Greenberg. Effect of heterogeneity and assumed mode of inheritance on lod scores. *Am. J. Med. Genet.*, 42(3):271–275, Feb 1992.

[29] A. S. Whittemore and J. Halpern. A class of tests for linkage using affected pedigree members. *Biometrics*, 50(1):118–127, Mar 1994.

[30] D. E. Weeks and K. Lange. The affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.*, 42(2):315–326, Feb 1988.

[31] J. M. Hall, M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, and M. C. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689, Dec 1990.

[32] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, and G. Micklem. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559):789–792, 1995.

[33] R. Arya, R. Duggirala, L. Almasy, D. L. Rainwater, M. C. Mahaney, S. Cole, T. D. Dyer, K. Williams, R. J. Leach, J. E. Hixson, J. W. MacCluer, P. O'Connell, M. P. Stern, and J. Blangero. Linkage of high-density lipoprotein-cholesterol concentrations to a locus on chromosome 9p in mexican americans. *Nat. Genet.*, 30(1):102–105, Jan 2002.

[34] M. A. Pericak-Vance, L. H. Yamaoka, C. S. Haynes, M. C. Speer, J. L. Haines, P. C. Gaskell, W. Y. Hung, C. M. Clark, A. L. Heyman, and J. A. Trofatter. Genetic linkage studies in alzheimer's disease families. *Exp. Neurol.*, 102(3):271–279, Dec 1988.

[35] J. Altmuller, L. J. Palmer, G. Fischer, H. Scherb, and M. Wjst. Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.*, 69(5):936–950, Nov 2001.

[36] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, Sep 1994.

[37] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am. J. Hum. Genet.*, 52(3):506–516, Mar 1993.

[38] A. Cao and R. Galanello. Beta-thalassemia. *Genet. Med.*, 12(2):61–76, Feb 2010.

[39] J. Ott. Statistical properties of the haplotype relative risk. *Genet. Epidemiol.*, 6(1):127–130, 1989.

[40] W. J. Ewens and R. S. Spielman. The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.*, 57(2):455–464, Aug 1995.

[41] L. L. Field, C. Fothergill-Payne, J. Bertrams, and M. P. Baur. Hla-dr effects in a large german iddm dataset. *Genet Epidemiol Suppl*, 1:323–328, 1986.

[42] E. P. Hong and J. W. Park. Sample size and statistical power calculation in genetic association studies. *Genomics Inform*, 10(2):117–122, Jun 2012.

[43] R. A. Fisher. On the interpretation of chisquare from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[44] WG Cochran. Some methods for strengthening the common chi-squared tests. *Biometrics*, 10(4):417–451, 1954.

[45] P Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[46] Benjamini Yoav and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B. 57 (1):289–300, 1995.

[47] O. A. Panagiotou and J. P. Ioannidis. What should the genome-wide significance threshold be? empirical replication of borderline genetic associations. *Int J Epidemiol*, 41(1):273–286, Feb 2012.

[48] No authors listed. The international hapmap project. *Nature*, 426(6968):789–796, Dec 2003.

[49] R. S. Spielman and W. J. Ewens. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.*, 62(2):450–458, Feb 1998.

[50] D. Rabinowitz and N. Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.*, 50(4):211–223, 2000.

[51] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, 30(1):97–101, Jan 2002.

[52] L. Kent, J. Emerton, V. Bhadravathi, E. Weisblatt, G. Pasco, L. R. Willatt, R. McMahon, and J. R. Yates. X-linked ichthyosis (steroid sulfatase deficiency) is associated with increased risk of attention deficit hyperactivity disorder, autism and social communication deficits. *J. Med. Genet.*, 45(8):519–524, Aug 2008.

[53] P. L. Roubertoux, P. V. Guillot, S. Mortaud, M. Pratte, M. Jamon, C. Cohen-Salmon, and S. Tordjman. Attack behaviors in mice: from factorial structure to quantitative trait loci mapping. *Eur. J. Pharmacol.*, 526(1-3):172–185, Dec 2005.

[54] L. B. Nicolas, W. Pinoteau, S. Papot, S. Routier, G. Guillaumet, and S. Mortaud. Aggressive behavior induced by the steroid sulfatase inhibitor coumate and by dheas in cba/h mice. *Brain Res.*, 922(2):216–222, Dec 2001.

[55] I. Le Roy, S. Mortaud, S. Tordjman, E. Donsez-Darcel, M. Carlier, H. Degrelle, and P. L. Roubertoux. *Behav. Genet.*, 29(2):131–136, Mar 1999.

[56] J. Semba. [glycine therapy of schizophrenia; its rationale and a review of clinical trials]. *Nihon Shinkei Seishin Yakurigaku Zasshi*, 18(3):71–80, Jun 1998.

[57] C. Meyer, R. Schmid, P. C. Scriba, and M. Wehling. Purification and partial sequencing of high-affinity progesterone-binding site(s) from porcine liver membranes. *Eur. J. Biochem.*, 239(3):726–731, Aug 1996.

[58] M. W. Wang, D. L. Crombie, J. S. Hayes, and R. B. Heap. Aberrant maternal behaviour in mice treated with a progesterone receptor antagonist during pregnancy. *J. Endocrinol.*, 145(2):371–377, May 1995.

[59] E. Dremencov, Y. Weizmann, N. Kinor, I. Gispan-Herman, and G. Yadid. Modulation of dopamine transmission by 5ht2c and 5ht3 receptors: a role in the antidepressant response. *Curr Drug Targets*, 7(2):165–175, Feb 2006.

[60] I. Gurevich, H. Tamir, V. Arango, A. J. Dwork, J. J. Mann, and C. Schmauss. Altered editing of serotonin 2c receptor pre-mrna in the prefrontal cortex of depressed suicide victims. *Neuron*, 34(3):349–356, Apr 2002.

[61] S. C. Pandey, L. Lumeng, and T. K. Li. Serotonin2c receptors and serotonin2c receptor-mediated phosphoinositide hydrolysis in the brain of alcohol-preferring and alcohol-nonpreferring rats. *Alcohol. Clin. Exp. Res.*, 20(6):1038–1042, Sep 1996.

[62] R. Trifunovic and S. Reilly. Medial parabrachial nucleus neurons modulate d-fenfluramine-induced anorexia through 5ht2c receptors. *Brain Res.*, 1067(1):170–176, Jan 2006.

[63] L. K. Heisler, M. A. Cowley, L. H. Tecott, W. Fan, M. J. Low, J. L. Smart, M. Rubinstein, J. B. Tatro, J. N. Marcus, H. Holstege, C. E. Lee, R. D. Cone, and J. K. Elmquist. Activation of central melanocortin pathways by fenfluramine. *Science*, 297(5581):609–611, Jul 2002.

[64] S. Kishore and S. Stamm. The snorna hbii-52 regulates alternative splicing of the serotonin receptor 2c. *Science*, 311(5758):230–232, Jan 2006.

[65] A. L. Lopez-Figueroa, C. S. Norton, M. O. Lopez-Figueroa, D. Armellini-Dodel, S. Burke, H. Akil, J. F. Lopez, and S. J. Watson. Serotonin 5-ht1a, 5-ht1b, and 5-ht2a receptor mrna expression in subjects with major depression, bipolar disorder, and schizophrenia. *Biol. Psychiatry*, 55(3):225–233, Feb 2004.

[66] T. Kato and N. Kato. Mitochondrial dysfunction in bipolar disorder. *Bipolar Disord*, 2(3 Pt 1):180–190, Sep 2000.

[67] D. Ben-Shachar. Mitochondrial dysfunction in schizophrenia: a possible linkage to dopamine. *J. Neurochem.*, 83(6):1241–1251, Dec 2002.

[68] R. S. Seelan and L. I. Grossman. Structural organization and promoter analysis of the bovine cytochrome c oxidase subunit viic gene. a functional role for yy1. *J. Biol. Chem.*, 272(15):10175–10181, Apr 1997.

[69] P. Korhonen, V. Huotari, H. Soininen, and A. Salminen. Glutamate-induced changes in the dna-binding complexes of transcription factor yy1 in cultured hippocampal and cerebellar granule cells. *Brain Res. Mol. Brain Res.*, 52(2):330–333, Dec 1997.

[70] W. R. Perlman, M. J. Webster, J. E. Kleinman, and C. S. Weickert. Reduced glucocorticoid and estrogen receptor alpha messenger ribonucleic acid levels in the amygdala of patients with major mental illness. *Biol. Psychiatry*, 56(11):844–852, Dec 2004.

[71] M. J. Webster, M. B. Knable, J. O'Grady, J. Orthmann, and C. S. Weickert. Regional specificity of brain glucocorticoid receptor mRNA alterations in subjects with schizophrenia and mood disorders. *Mol. Psychiatry*, 7(9):985–994, 2002.

[72] Kaabi B., Gelernter J., Woods S. W., Goddard A., Page G. P., and Elston R. C. Genome scan for loci predisposing to anxiety disorders using a novel multivariate approach: strong evidence for a chromosome 4 risk locus. *Am. J. Hum. Genet.*, 78(4):543–553, Apr 2006.

[73] F. Gachon, P. Fonjallaz, F. Damiola, P. Gos, T. Kodama, J. Zakany, D. Duboule, B. Petit, M. Tafti, and U. Schibler. The loss of circadian par bzip transcription factors results in epilepsy. *Genes Dev.*, 18(12):1397–1412, Jun 2004.

[74] G. J. Ho, R. Drego, E. Hakimian, and E. Masliah. Mechanisms of cell signaling and inflammation in alzheimer's disease. *Curr Drug Targets Inflamm Allergy*, 4(2):247–256, Apr 2005.

[75] C. L. Hunter, D. Bachman, and A. C. Granholm. Minocycline prevents cholinergic loss in a mouse model of down's syndrome. *Ann. Neurol.*, 56(5):675–688, Nov 2004.

[76] A. F. Arnsten. Adrenergic targets for the treatment of cognitive deficits in schizophrenia. *Psychopharmacology (Berl.)*, 174(1):25–31, Jun 2004.

[77] J. Lahdesmaki, J. Sallinen, E. MacDonald, and M. Scheinin. Alpha2a-adrenoceptors are important modulators of the effects of D-amphetamine on startle reactivity and brain monoamines. *Neuropsychopharmacology*, 29(7):1282–1293, Jul 2004.

[78] J. N. Wood, J. P. Boorman, K. Okuse, and M. D. Baker. Voltage-gated sodium channels and pain pathways. *J. Neurobiol.*, 61(1):55–71, Oct 2004.

[79] M. P. Boks, M. Hoogendoorn, B. J. Jungerius, S. C. Bakker, I. E. Sommer, R. J. Sinke, R. A. Ophoff, and R. S. Kahn. Do mood symptoms subdivide the schizophrenia phenotype? Association of the GMP6A gene with a depression subgroup. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 147B(6):707–711, Sep 2008.

[80] C. R. Quilter, C. L. Gilbert, G. L. Oliver, O. Jafer, R. A. Furlong, S. C. Blott, A. E. Wilson, C. A. Sargent, A. Mileham, and N. A. Affara. Gene expression profiling in porcine maternal infanticide: a model for puerperal psychosis. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 147B(7):1126–1137, Oct 2008.

[81] R. Kellner, M. T. Buckman, G. A. Fava, and D. Pathak. Hyperprolactinemia, distress, and hostility. *Am J Psychiatry*, 141(6):759–763, Jun 1984.

[82] P. Fitzgerald and T. G. Dinan. Prolactin and dopamine: what is the connection? a review article. *J. Psychopharmacol. (Oxford)*, 22(2 Suppl):12–19, Mar 2008.

[83] A. K. Ragnauth, N. Devidze, V. Moy, K. Finley, A. Goodwillie, L. M. Kow, L. J. Muglia, and D. W. Pfaff. Female oxytocin gene-knockout mice, in a semi-natural environment, display exaggerated aggressive behavior. *Genes Brain Behav.*, 4(4):229–239, Jun 2005.

[84] J. Frasor and G. Gibori. Prolactin regulation of estrogen receptor expression. *Trends Endocrinol. Metab.*, 14(3):118–123, Apr 2003.

[85] E. P. Noble. D2 dopamine receptor gene in psychiatric and neurologic disorders and its phenotypes. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 116B(1):103–125, Jan 2003.

[86] C. Rosin, S. Colombo, A. A. Calver, T. E. Bates, and S. D. Skaper. Dopamine d2 and d3 receptor agonists limit oligodendrocyte injury caused by glutamate oxidative stress and oxygen/glucose deprivation. *Glia*, 52(4):336–343, Dec 2005.

[87] N. A. Uranova, V. M. Vostrikov, O. V. Vikhreva, I. S. Zimina, N. S. Kolomeets, and D. D. Orlovskaya. The role of oligodendrocyte pathology in schizophrenia. *Int. J. Neuropsychopharmacol.*, 10(4):537–545, Aug 2007.

[88] M. J. Kas, R. van den Bos, A. M. Baars, M. Lubbers, H. M. Lesscher, J. J. Hillebrand, A. G. Schuller, J. E. Pintar, and B. M. Spruijt. Mu-opioid receptor knockout mice show diminished food-anticipatory activity. *Eur. J. Neurosci.*, 20(6):1624–1632, Sep 2004.

[89] E. Dremencov, Y. Weizmann, N. Kinor, I. Gispan-Herman, and G. Yadid. Modulation of dopamine transmission by 5ht2c and 5ht3 receptors: a role in the antidepressant response. *Curr Drug Targets*, 7(2):165–175, Feb 2006.

[90] N. Coyle, I. Jones, E. Robertson, C. Lendon, and N. Craddock. Variation at the serotonin transporter gene influences susceptibility to bipolar affective puerperal psychosis. *Lancet*, 356(9240):1490–1491, Oct 2000.

[91] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. Evans, H. T. Leung, J. L. Marchini, A. P. Morris, C. C. Spencer, M. D. Tobin, L. R. Cardon, D. G. Clayton, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, J. A. Todd, W. H. Ouwehand, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, N. Craddock, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, D. T. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, P. R. Burton, R. J. Dixon, M. Mangino, S. Suzanne, M. D. Tobin, J. R. Thompson, N. J. Samani, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi,

S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Mathew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, D. G. Clayton, G. M. Lathrop, J. Connell, A. Dominczak, N. J. Samani, C. A. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hider, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. Symmmons, W. Thomson, J. Worthington, D. G. Clayton, D. B. Dunger, S. Nutland, H. E. Stevens, N. M. Walker, B. Widmer, J. A. Todd, T. A. Frayling, R. M. Freathy, H. Lango, J. R. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. I. McCarthy, M. Newport, G. Sirugo, E. Lyons, F. Vannberg, A. V. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. Gough, S. Seal, M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, D. P. Kwiatowski, S. J. Bumpstead, A. Chaney, K. Downes, M. J. Ghori, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter, R. Ravindrarajah, P. Whittaker, C. Widden, D. Withers, P. Deloukas, H. T. Leung, S. Nutland, H. E. Stevens, N. M. Walker, J. A. Todd, D. Easton, D. G. Clayton, P. R. Burton, M. D. Tobin, J. C. Barrett, D. Evans, A. P. Morris, L. R. Cardon, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrimsdottir, B. N. Howie, J. L. Marchini, C. C. Spencer, Z. Su, Y. Y. Teo, D. Vukcevic, P. Donnelly, D. Bentley, M. A. Brown, L. R. Gordon, M. Caulfield, D. G. Clayton, A. Compston, N. Craddock, P. Deloukas, P. Donnelly, M. Farrall, S. C. Gough, A. S. Hall, A. T. Hattersley, A. V. Hill, D. P. Kwiatkowski, C. Mathew, M. I. McCarthy, W. H. Ouwehand, M. Parkes, M. Pembrey, N. Rahman, N. J. Samani, M. R. Stratton, J. A. Todd, and J. Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.

[92] S. Begni, S. Moraschi, S. Bignotti, F. Fumagalli, L. Rillosi, J. Perez, and M. Gennarelli. Association between the g1001c polymorphism in the grin1 gene promoter region and schizophrenia. *Biol. Psychiatry*, 53(7):617–619, Apr 2003.

[93] S. Y. Cherlyn, P. S. Woon, J. J. Liu, W. Y. Ong, G. C. Tsai, and K. Sim. Genetic association studies of glutamate, gaba and related genes in schizophrenia and bipolar disorder: a decade of advance. *Neurosci Biobehav Rev*, 34(6):958–977, May 2010.

[94] A. Sidhu and H. B. Niznik. Coupling of dopamine receptor subtypes to multiple and diverse g proteins. *Int. J. Dev. Neurosci.*, 18(7):669–677, Nov 2000.

[95] Eun Kyung Kim and Eui-Ju Choi. Pathological roles of mapk signaling pathways in human

diseases. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1802(4):396 – 405, 2010.

[96] J. C. Sousa, C. Grandela, J. Fernandez-Ruiz, R. de Miguel, L. de Sousa, A. I. Magalhaes, M. J. Saraiva, N. Sousa, and J. A. Palha. Transthyretin is involved in depression-like behaviour and exploratory activity. *J. Neurochem.*, 88(5):1052–1058, Mar 2004.

[97] N. Ahmadiyeh, G. A. Churchill, K. Shimomura, L. C. Solberg, J. S. Takahashi, and E. E. Redei. X-linked and lineage-dependent inheritance of coping responses to stress. *Mamm. Genome*, 14(11):748–757, Nov 2003.

[98] A. M. Ramos, R. P. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z. L. Hu, H. H. Kerstens, A. S. Law, H. J. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. Groenen. Design of a high density snp genotyping assay in the pig using snps identified and characterized by next generation sequencing technology. *PLoS ONE*, 4(8):e6524, Aug 2009.

[99] D. M. Dick, T. Foroud, H. J. Edenberg, M. Miller, E. Bowman, N. L. Rau, J. R. DePaulo, M. McInnis, E. Gershon, F. McMahon, J. P. Rice, L. J. Bierut, T. Reich, and J. Nurnberger. Apparent replication of suggestive linkage on chromosome 16 in the NIMH genetics initiative bipolar pedigrees. *Am. J. Med. Genet.*, 114(4):407–412, May 2002.

[100] H. J. Edenberg, T. Foroud, P. M. Conneally, J. J. Sorbel, K. Carr, C. Crose, C. Willig, J. Zhao, M. Miller, E. Bowman, A. Mayeda, N. L. Rau, C. Smiley, J. P. Rice, A. Goate, T. Reich, O. C. Stine, F. McMahon, J. R. DePaulo, D. Meyers, S. D. Detera-Wadleigh, L. R. Goldin, E. S. Gershon, M. C. Blehar, and J. I. Nurnberger. Initial genomic scan of the nimh genetics initiative bipolar pedigrees: chromosomes 3, 5, 15, 16, 17, and 22. *Am. J. Med. Genet.*, 74(3):238–246, May 1997.

[101] N. M. Williams, I. Zaharieva, A. Martin, K. Langley, K. Mantripragada, R. Fossdal, H. Stefansson, K. Stefansson, P. Magnusson, O. O. Gudmundsson, O. Gustafsson, P. Holmans, M. J. Owen, M. O'Donovan, and A. Thapar. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet*, 376(9750):1401–1408, Oct 2010.

[102] Y. Liu, D. H. Blackwood, S. Caesar, E. J. de Geus, A. Farmer, M. A. Ferreira, I. N. Ferrier, C. Fraser, K. Gordon-Smith, E. K. Green, D. Grozeva, H. M. Gurling, M. L. Hamshere, P. Heutink, P. A. Holmans, W. J. Hoogendijk, J. J. Hottenga, L. Jones, I. R. Jones, G. Kirov, D. Lin, P. McGuffin, V. Moskvina, W. A. Nolen, R. H. Perlis, D. Posthuma, E. M. Scolnick, A. B. Smit, J. H. Smit, J. W. Smoller, D. St Clair, R. van Dyck, M. Verhage, G. Willemsen,

A. H. Young, T. Zandbelt, D. I. Boomsma, N. Craddock, M. C. O'Donovan, M. J. Owen, B. W. Penninx, S. Purcell, P. Sklar, and P. F. Sullivan. *Mol. Psychiatry*, 16(1):2–4, Jan 2011.

[103] R. Holt, G. Barnby, E. Maestrini, E. Bacchelli, D. Brocklebank, I. Sousa, E. J. Mulder, K. Kantojarvi, I. Jarvela, S. M. Klauck, F. Poustka, A. J. Bailey, A. P. Monaco, R. Alen, E. Bacchelli, A. Bailey, G. Baird, A. Battaglia, C. Betancur, A. De Bildt, F. Blasi, S. Bolte, P. Bolton, T. Bourgeron, K. Br?ndum-Nielsen, S. Carone, P. Chaste, A. Chiocchetti, E. Duketis, C. Durand, H. Van Engeland, P. Farrar, S. Feineis-Matthews, B. Felder, K. Francis, J. Fremolle, C. Gillberg, C. Gillberg, H. Goubran-Botros, D. Haracopos, E. Herbrecht, R. Holt, G. Honeyman, J. Honold, R. Houben, A. Hutchison, R. Igliozzi, T. Isager, I. Jarvela, M. Johansson, M. De Jonge, S. M. Klauck, A. Koivisto, H. Komu, M. Leboyer, A. Le Couteur, J. Lowen, E. Maestrini, C. Mantoulan, J. Melke, H. McConachie, R. Minderaa, A. Monaco, E. Mulder, T. Nieminen-von Wendt, I. Nummela, G. Nygren, G. Pakalapati, K. Papanikolaou, J. Parr, B. Parrini, L. Pederson, L. Pellicano, C. Pienkowski, J. Ponsford, A. Poustka, F. Poustka, M. Rastam, K. Rehnstrom, K. Renshaw, B. Roge, D. Ruehl, M. Rutter, S. Sarenius, G. Schmotzer, C. Schuster, H. Anckarsater, R. Tancredi, M. Tauber, J. Tsiantis, N. Uhlig, R. Vanhala, S. Wallace, L. Von Wendt, K. Wittemeyer, and T. Ylisaukko-oja. Linkage and candidate gene studies of autism spectrum disorders in European populations. *Eur. J. Hum. Genet.*, 18(9):1013–1019, Sep 2010.

[104] D. M. Dick, F. Aliev, R. F. Krueger, A. Edwards, A. Agrawal, M. Lynskey, P. Lin, M. Schuckit, V. Hesselbrock, J. Nurnberger, L. Almasy, B. Porjesz, H. J. Edenberg, K. Bucholz, J. Kramer, S. Kuperman, and L. Bierut. Genome-wide association study of conduct disorder symptomatology. *Mol. Psychiatry*, 16(8):800–808, Aug 2011.

[105] M. C. O'Donovan, N. Craddock, N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. Spencer, B. Howie, H. T. Leung, A. M. Hartmann, H. J. Moller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. M. Williams, I. Giegling, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, P. V. Gejman, S. Cichon, M. M. Nothen, M. Gill, A. Corvin, D. Rujescu, G. Kirov, M. J. Owen, N. G. Buccola, B. J. Mowry, R. Freedman, F. Amin, D. W. Black, J. M. Silverman, W. F. Byerley, and C. R. Cloninger. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat. Genet.*, 40(9):1053–1055, Sep 2008.

[106] Ute Süsens, Irm Hermans-Borgmeyer, Jens Urny, and H Chica Schaller. Characterisation and differential expression of two very closely related g-protein-coupled receptors, gpr139 and gpr142, in mouse tissue and during mouse development. *Neuropharmacology*, 50(4):512–520, 2006.

[107] D. E. Gloriam, H. B. Schioth, and R. Fredriksson. Nine new human rhodopsin family g-protein coupled receptors: identification, sequence characterisation and evolutionary relationship. *Biochim. Biophys. Acta*, 1722(3):235–246, Apr 2005.

[108] Ian Jones, Marian Hamshere, Jeanne-Marrie Nangle, Philip Bennett, Elaine Green, Jess Heron, Ricardo Segurado, David Lambert, Peter Holmans, Aiden Corvin, Mike Owen, Lisa Jones, Michael Gill, and Nick Craddock. Bipolar affective puerperal psychosis: Genome-wide significant evidence for linkage to chromosome 16. *Am J Psychiatry*, 164:1099–1104, 2007.

[109] H. Ewald, T. Flint, T. A. Kruse, and O. Mors. A genome-wide scan shows significant linkage between bipolar disorder and chromosome 12q24.3 and suggestive linkage to chromosomes 1p22-21, 4p16, 6q14-22, 10q26 and 16p13.3. *Mol. Psychiatry*, 7(7):734–744, 2002.

[110] Hiroki Shibata, Duong P. Huynh, and Stefan-M. Pulst. A novel protein with rna-binding motifs interacts with ataxin-2. *Human Molecular Genetics*, 9(9):1303–1313, 2000.

[111] K. Aberg, D. E. Adkins, J. Bukszar, B. T. Webb, S. N. Caroff, D. D. Miller, J. Sebat, S. Stroup, A. H. Fanous, V. I. Vladimirov, J. L. McClay, J. A. Lieberman, P. F. Sullivan, and E. J. van den Oord. Genomewide association study of movement-related adverse antipsychotic effects. *Biol. Psychiatry*, 67(3):279–282, Feb 2010.

[112] Christa Lese Martin, Jacqueline A. Duvall, Yesim Ilkin, Jason S. Simon, M. Gladys Arreaza, Kristin Wilkes, Ana Alvarez-Retuerto, Amy Whichello, Cynthia M. Powell, Kathleen Rao, Edwin Cook, and Daniel H. Geschwind. Cytogenetic and molecular characterization of a2bp1/fox1 as a candidate gene for autism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 144B(7):869–876, 2007.

[113] J. Elia, X. Gai, H. M. Xie, J. C. Perin, E. Geiger, J. T. Glessner, M. D'arcy, R. deBerardinis, E. Frackelton, C. Kim, F. Lantieri, B. M. Muganga, L. Wang, T. Takeda, E. F. Rappaport, S. F. Grant, W. Berrettini, M. Devoto, T. H. Shaikh, H. Hakonarson, and P. S. White. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol. Psychiatry*, 15(6):637–646, Jun 2010.

[114] R. J. Anney, J. Lasky-Su, C. O'Dushlaine, E. Kenny, B. M. Neale, A. Mulligan, B. Franke, K. Zhou, W. Chen, H. Christiansen, A. Arias-Vasquez, T. Banaschewski, J. Buitelaar, R. Ebstein, A. Miranda, F. Mulas, R. D. Oades, H. Roeyers, A. Rothenberger, J. Sergeant, E. Sonuga-Barke, H. Steinhausen, P. Asherson, S. V. Faraone, and M. Gill. Conduct disorder and adhd: evaluation of conduct problems as a categorical and quantitative trait in the international multicentre adhd genetics study. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 147B(8):1369–1378, Dec 2008.

[115] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, and S. T. Sherry. The ncbi dbgap database of genotypes and phenotypes. *Nat. Genet.*, 39(10):1181–1186, Oct 2007.

[116] Gabrielle Barnby, Aaron Abbott, Nuala Sykes, Andrew Morris, Daniel E. Weeks, Richard Mott, Janine Lamb, Anthony J. Bailey, and Anthony P. Monaco. Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: Evidence of association at grin2a and abat. *The American Journal of Human Genetics*, 76(6):950 – 966, 2005.

[117] Hon-Chung Fung, Sonja Scholz, Mar Matarin, Javier Simón-Sánchez, Dena Hernandez, Angela Britton, J Raphael Gibbs, Carl Langefeld, Matt L Stiegert, Jennifer Schymick, Michael S Okun, Ronald J Mandel, Hubert H Fernandez, Kelly D Foote, Ramón L Rodríguez, Elizabeth Peckham, Fabienne Wavrant De Vrieze, Katrina Gwinn-Hardy, John A Hardy, and Andrew Singleton. Genome-wide genotyping in parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, 5(11):911–916, November 2006.

[118] E. J. van den Oord, P. H. Kuo, A. M. Hartmann, B. T. Webb, H. J. Moller, J. M. Hettema, I. Giegling, J. Bukszar, and D. Rujescu. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Arch. Gen. Psychiatry*, 65(9):1062–1071, Sep 2008.

[119] E. T. Cirulli, D. Kasperavici?te, D. K. Attix, A. C. Need, D. Ge, G. Gibson, and D. B. Goldstein. Common genetic variation and performance on standardized cognitive tests. *Eur. J. Hum. Genet.*, 18(7):815–820, Jul 2010.

[120] H. Wei, M. Malik, A. M. Sheikh, G. Merz, W. Ted Brown, and X. Li. Abnormal cell properties and down-regulated FAK-Src complex signaling in B lymphoblasts of autistic subjects. *Am. J. Pathol.*, 179(1):66–74, Jul 2011.

[121] Barel O., Shalev S. A., Ofir R., Cohen A., Zlotogora J., Shorer Z., Mazor G., Finer G., Khateeb S., Zilberberg N., and Birk O. S. Maternally inherited Birk Barel mental retardation dysmorphism syndrome caused by a mutation in the genomically imprinted potassium channel KCNK9. *Am. J. Hum. Genet.*, 83(2):193–199, Aug 2008.

[122] P. B. Mahon, J. L. Payne, D. F. MacKinnon, F. M. Mondimore, F. S. Goes, B. Schweizer, D. Jancic, W. H. Coryell, P. A. Holmans, J. Shi, J. A. Knowles, W. A. Scheftner, M. M. Weissman, D. F. Levinson, J. R. DePaulo, P. P. Zandi, J. B. Potash, M. M. Weissman, J. K. Knowles, G. S. Zubenko, W. N. Zubenko, J. R. DePaulo, M. G. McInnis, D. MacKinnon, D. F. Levinson,

M. M. Gladis, K. Murphy-Eberenz, P. Holmans, R. R. Crowe, W. H. Coryell, W. A. Scheftner, J. Nurnberger, M. Miller, E. Bowman, T. Reich, A. Goate, J. Rice, J. R. DePaulo, S. Simpson, C. Stine, E. Gershon, D. Kazuba, E. Maxwell, J. Nurnberger, M. J. Miller, E. S. Bowman, N. L. Rau, P. R. Moe, N. Samavedy, R. El-Mallakh, H. Manji, D. A. Glitz, E. T. Meyer, C. Smiley, T. Foroud, L. Flury, D. M. Dick, H. Edenberg, J. Rice, T. Reich, A. Goate, L. Bierut, M. McInnis, J. R. DePaulo, D. F. MacKinnon, F. M. Mondimore, J. B. Potash, P. P. Zandi, D. Avramopoulos, J. Payne, W. Berrettini, W. Byerley, M. Vawter, W. Coryell, R. Crowe, E. Gershon, J. Badner, F. McMahon, C. Liu, A. Sanders, M. Caserta, S. Dinwiddie, T. Nguyen, D. Harakal, J. Kelsoe, R. McKinney, W. Scheftner, H. M. Kravitz, D. Marta, A. Vaughn-Brown, L. Bederow, F. J. McMahon, L. Kassem, S. Detera-Wadleigh, L. Austin, D. L. Murphy, P. V. Gejman, A. R. Sanders, F. Amin, N. Buccola, W. Byerley, C. R. Cloninger, R. Crowe, D. Black, R. Freedman, D. Levinson, B. Mowry, J. Silverman, J. R. Kelsoe, T. A. Greenwood, C. Nievergelt, N. Schork, E. N. Smith, C. Bloss, J. Nurnberger, H. J. Edenberg, T. Foroud, E. Gershon, C. Liu, J. A. Badner, W. A. Scheftner, W. B. Lawson, E. A. Nwulia, M. Hipolito, W. Coryell, J. Rice, W. Byerley, F. McMahon, T. G. Schulze, W. Berrettini, J. B. Potash, P. P. Zandi, P. B. Mahon, M. G. McInnis, S. Zollner, D. Craig, and S. Szelinger. Genome-wide linkage and follow-up association study of postpartum mood symptoms. *Am J Psychiatry*, 166(11):1229–1237, Nov 2009.

[123] D J Epstein, K J Vogan, D G Trasler, and P Gros. A mutation within intron 3 of the pax-3 gene produces aberrantly spliced mrna transcripts in the splotch (sp) mouse mutant. *Proceedings of the National Academy of Sciences*, 90(2):532–536, 1993.

[124] Asher J. H., Harrison R. W., Morell R., Carey M. L., and Friedman T. B. Effects of pax3 modifier genes on craniofacial morphology, pigmentation, and viability: a murine model of waardenburg syndrome variation. *Genomics*, 34(3):285–298, Jun 1996.

[125] Wilkinson D. G. Multiple roles of eph receptors and ephrins in neural development. *Nat. Rev. Neurosci.*, 2(3):155–164, Mar 2001.

[126] L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. D. Sloan, M. J. Huentelman, D. W. Craig, B. M. Dechairo, S. G. Potkin, C. R. Jack, M. W. Weiner, A. J. Saykin, M. Weiner, P. Aisen, M. Weiner, P. Aisen, R. Petersen, C. R. Jack, W. Jagust, J. Trojanowki, A. W. Toga, L. Beckett, R. C. Green, A. Gamst, A. J. Saykin, J. Morris, W. Z. Potter, R. C. Green, T. Montine, R. Petersen, P. Aisen, A. Gamst, R. G. Thomas, M. Donohue, S. Walter, C. R. Jack, A. Dale, M. Bernstein, J. Felmlee, N. Fox, P. Thompson, N. Schuf, G. Alexander, C. DeCarli, W. Jagust, D. Bandy, R. A. Koeppe, N. Foster, E. M. Reiman, K. Chen, C. Mathis, J. Morris, N. J. Cairns, L. Taylor-Reinwald, J. Trojanowki,

L. Shaw, V. M. Lee, M. Korecka, A. W. Toga, K. Crawford, S. Neu, L. Beckett, D. Harvey, A. Gamst, J. Kornak, A. J. Saykin, T. M. Foroud, S. Potkin, L. Shen, Z. Kachaturian, R. Frank, P. J. Snyder, S. Molchan, J. Kaye, S. Dolen, J. Quinn, L. Schneider, S. Pawluczyk, B. M. Spann, J. Brewer, H. Vanderswag, J. L. Heidebrink, J. L. Lord, R. Petersen, K. Johnson, R. S. Doody, J. Villanueva-Meyer, M. Chowdhury, Y. Stern, L. S. Honig, K. L. Bell, J. C. Morris, M. A. Mintun, S. Schneider, D. Marson, R. Griffith, D. Clark, H. Grossman, C. Tang, G. Marzloff, L. deToledo Morrell, R. C. Shah, R. Duara, D. Varon, P. Roberts, M. S. Albert, N. Kozauer, M. Zerrate, H. Rusinek, M. J. de Leon, S. M. De Santi, P. M. Doraiswamy, J. R. Petrella, M. Aiello, S. Arnold, J. H. Karlawish, D. Wolk, C. D. Smith, C. A. Given, P. Hardy, O. L. Lopez, M. Oakley, D. M. Simpson, M. S. Ismail, C. Brand, J. Richard, R. A. Mulnard, G. Thai, C. Mc-Adams-Ortiz, R. Diaz-Arrastia, K. Martin-Cook, M. DeVous, A. I. Levey, J. J. Lah, J. S. Cellar, J. M. Burns, H. S. Anderson, M. M. Laubinger, L. Apostolova, D. H. Silverman, P. H. Lu, N. R. Graff-Radford, F. Parfitt, H. Johnson, M. Farlow, S. Herring, A. M. Hake, C. H. van Dyck, M. G. MacAvoy, A. L. Benincasa, H. Chertkow, H. Bergman, C. Hosein, S. Black, B. Stefanovic, C. Caldwell, G. Y. Hsiung, H. Feldman, M. Assaly, A. Kertesz, J. Rogers, D. Trost, C. Bernick, D. Munic, C. K. Wu, N. Johnson, M. Mesulam, C. Sadowsky, W. Martinez, T. Villena, R. S. Turner, K. Johnson, B. Reynolds, R. A. Sperling, D. M. Rentz, K. A. Johnson, A. Rosen, J. Tinklenberg, W. Ashford, M. Sabbagh, D. Connor, S. Jacobson, R. Killiany, A. Norbash, A. Nair, T. O. Obisesan, A. Jayam-Trouth, P. Wang, A. Lerner, L. Hudson, P. Ogrocki, C. DeCarli, E. Fletcher, O. Carmichael, S. Kittur, M. Borrie, T. Y. Lee, R. Bartha, S. Johnson, S. Asthana, C. M. Carlsson, S. G. Potkin, A. Preda, D. Nguyen, P. Tariot, A. Fleisher, S. Reeder, V. Bates, H. Capote, M. Rainka, B. A. Hendin, D. W. Scharre, M. Kataki, E. A. Zimmerman, D. Celmins, A. D. Brown, G. Pearlson, K. Blank, K. Anderson, A. J. Saykin, R. B. Santulli, J. Englert, J. D. Williamson, K. M. Sink, F. Watkins, B. R. Ott, E. Stopa, G. Tremont, S. Salloway, P. Malloy, S. Correia, H. J. Rosen, B. L. Miller, J. Mintzer, C. F. Longmire, and K. Spicer. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohort. *Neuroimage*, 53(3):1051–1063, Nov 2010.

[127] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, Feb 2001.

[128] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

[129] M. B. Miller and Y. W. Tang. Basic concepts of microarrays and potential applications in clinical microbiology. *Clin. Microbiol. Rev.*, 22(4):611–633, Oct 2009.

[130] Gordon K Smyth and Terry Speed. Normalization of cdna microarray data. *Methods*, 31(4):265 – 273, 2003. Candidate Genes from {DNA} Array Screens: application to neuroscience.

[131] T. Goldmann and J. S. Gonzalez. Dna-printing: utilization of a standard inkjet printer for the transfer of nucleic acids to solid supports. *J. Biochem. Biophys. Methods*, 42(3):105–110, Mar 2000.

[132] M. J. Heller. Dna microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, 4:129–153, 2002.

[133] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker. The microarray quality c ontrol (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, 24(9):1151–1161, Sep 2006.

[134] Leonid Kruglyak. The road to genome-wide association studies. *Nat Rev Genet*, 9(4):314–318, April 2008.

[135] M. Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, 9(6):477–485, Jun 2008.

[136] T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.*, 37(13):4181–4193, Jul 2009.

[137] M. A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagace, B. Neale, K. S. Lo, P. Schumm, L. Torkvist, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, G. Lettre, A. Franke, M. D'Amato, D. P. McGovern, J. H. Cho, J. D. Rioux, R. J. Xavier, M. J. Daly, S. R. Brant, J. H. Cho, R. H. Duerr, D. P. McGovern, J. D. Rioux, M. S. Silverberg, M. Parkes, J. Lee, H. Zhang, F. Bredin, T. Ahmad, J. Satsangi, E. Nimmo, H. Drummond, C. Lees, J. Mansfield, C. G. Mathew, N. Prescott, K. Harrison, J. Sanderson, W. Newman, A. Phillips, C. Mowat, C. Edwards, D. C. Wilson, J. Barrett, C. Anderson, E. Gray, S. Edkins, R. K. Russell, P. Henderson, T. Ahmad, C. A. Anderson, V. Annese, R. N. Baldassano, T. Balschun, M. Barclay, J. C. Barrett, T. M. Bayless, J. C. Bis, S. Brand, S. R. Brant, S. Bumpstead, C. Buning, J. H. Cho, A. Cohen, J. F. Colombel, M. Cottone, M. D'Amato, R. D'Inca, M. J. Daly, T. Denson, M. Dubinsky, R. H. Duerr, C. Edwards, D. Ellinghaus, T. Florin, D. Franchimont, A. Franke, R. Gearry, M. Georges, J. Glas, A. Van Gossum, A. M. Griffiths, S. L. Guthery, H. Hakonarson, T. Haritunians, J. P. Hugot, D. J. de Jong, L. Jostins, S. Kugathasan, G. Kullack-Ublick, A. Latiano, D. Laukens, I. Lawrance, J. Lee, C. W. Lees, M. Lemann, A. Levine, C. Libioulle, E. Louis, J. C. Mansfield, C. G. Mathew, D. P. McGovern, M. Mitrovic, G. W. Montgomery, C. Mowat, W. Newman, O. Palmieri, J. Panes, M. Parkes, A. Phillips, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, D. D. Proctor, G. L. Radford-Smith, M. Regueiro, J. D. Rioux, R. Roberts, J. I. Rotter, P. Rutgeerts, J. Sanderson, M. Sans, J. Satsangi, S. Schreiber, P. Schumm, F. Seibold, Y. Sharma, M. S. Silverberg, L. A. Simms, A. H. Steinhart, S. R. Targan, K. D. Taylor, L. Torkvist, S. Vermeire, J. Halfvarson, H. W. Verspaget, M. De Vos, T. Walters, K. Wang, R. K. Weersma, D. Whiteman, and C. Wijmenga. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, 43(11):1066–1073, Oct 2011.

[138] M. Beaudoin, P. Goyette, G. Boucher, K. S. Lo, M. A. Rivas, C. Stevens, A. Alikashani, M. Ladouceur, D. Ellinghaus, L. Torkvist, G. Goel, C. Lagace, V. Annese, A. Bitton, J. Begun, S. R. Brant, F. Bresso, J. H. Cho, R. H. Duerr, J. Halfvarson, D. P. McGovern, G. Radford-Smith, S. Schreiber, P. L. Schumm, Y. Sharma, M. S. Silverberg, R. K. Weersma, M. D'Amato, S. Vermeire, A. Franke, G. Lettre, R. J. Xavier, M. J. Daly, J. D. Rioux, G. Aumais, E. J. Bernard, A. Bitton, A. Cohen, C. Deslandres, R. Lahaie, P. Pare, J. D. Rioux, S. R. Brant, J. H. Cho, R. H. Duerr, D. P. McGovern, J. D. Rioux, M. S. Silverberg, T. Ahmad, C. A. Anderson,

V. Annese, R. N. Baldassano, T. Balschun, M. Barclay, J. C. Barrett, T. M. Bayless, J. C. Bis, S. Brand, S. R. Brant, S. Bumpstead, C. Buning, J. H. Cho, A. Cohen, J. F. Colombel, M. Cottone, M. D'Amato, R. D'Inca, M. J. Daly, T. Denson, M. Dubinsky, R. H. Duerr, C. Edwards, D. Ellinghaus, T. Florin, D. Franchimont, A. Franke, R. Gearry, M. Georges, J. Glas, A. Van Gossum, A. M. Griffiths, S. L. Guthery, H. Hakonarson, T. Haritunians, J. P. Hugot, D. J. de Jong, L. Jostins, S. Kugathasan, G. Kullak-Ublick, A. Latiano, D. Laukens, I. Lawrance, J. Lee, C. W. Lees, M. Lemann, A. Levine, C. Libioulle, E. Louis, J. C. Mansfield, C. G. Mathew, D. P. McGovern, M. Mitrovic, G. W. Montgomery, C. Mowat, W. Newman, O. Palmieri, J. Panes, M. Parkes, A. Phillips, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, D. D. Proctor, G. L. Radford-Smith, M. Regueiro, J. D. Rioux, R. Roberts, J. I. Rotter, P. Rutgeerts, J. Sanderson, M. Sans, J. Satsangi, S. Schreiber, P. Schumm, F. Seibold, Y. Sharma, M. S. Silverberg, L. A. Simms, A. Steinhart, S. R. Targan, K. D. Taylor, L. Torkvist, S. Vermeire, J. Halfvarson, H. W. Verspaget, M. De Vos, T. Walters, K. Wang, R. K. Weersma, D. Whiteman, and C. Wijmenga. Deep resequencing of gwas loci identifies rare variants in card9, il23r and rnf186 that are associated with ulcerative colitis. *PLoS Genet.*, 9(9):e1003723, 2013.

[139] S. E. Calvo, E. J. Tucker, A. G. Compton, D. M. Kirby, G. Crawford, N. P. Burtt, M. Rivas, C. Guiducci, D. L. Bruno, O. A. Goldberger, M. C. Redman, E. Wiltshire, C. J. Wilson, D. Altshuler, S. B. Gabriel, M. J. Daly, D. R. Thorburn, and V. K. Mootha. High-throughput, pooled sequencing identifies mutations in nubpl and foxred1 in human complex i deficiency. *Nat. Genet.*, 42(10):851–858, Oct 2010.

[140] S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd. Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–389, Apr 2009.

[141] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8, Jan 2016.

[142] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, May 2016.

[143] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang,

M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[144] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[145] W-K Hon; T-W Lam; K. Sadakane; W-K Sung; S-M Yiu. A space and time efficient algorithm for constructing compressed suffix arrays. *Algorithmica*, 48(1):23–36, March 2007.

[146] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. "substantial biases in ultra-short read data sets from high-throughput dna sequencing". *Nucleic Acids Res.*, 36(16):e105, Sep 2008.

[147] L. J. Manley, D. Ma, and S. S. Levine. Monitoring error rates in illumina sequencing. *J Biomol Tech*, 27(4):125–128, 12 2016.

[148] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella,

D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data.

[149] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

[150] L. B. Schook, J. E. Beever, J. Rogers, S. Humphray, A. Archibald, P. Chardon, D. Milan, G. Rohrer, and K. Eversole. Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome. *Comp. Funct. Genomics*, 6(4):251–255, 2005.

[151] H. Shizuya, B. Birren, U. J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. U.S.A.*, 89(18):8794–8797, Sep 1992.

[152] B. L. Aken, P. Achuthan, W. Akanni, M. R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, T. Juettemann, S. Keenan, M. R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D. N. Murphy, R. Nag, V. Newman, M. Nuhn, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S. P. Wilder, A. Zadissa, M. Kostadima, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, F. Cunningham, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2017. *Nucleic Acids Res.*, 45(D1):D635–D642, Jan 2017.

[153] Green P. Smit AFA, Huble R. Repeatmasker open-4.0, 2013-2015. <http://www.repeatmasker.org>.

[154] A. Morgulis, E. M. Gertz, A. A. Schaffer, and R. Agarwala. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, 13(5):1028–1040, Jun 2006.

[155] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94, Apr 1997.

[156] T. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 46(5):2699, Mar 2018.

[157] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Res.*, 33(Database issue):D34–38, Jan 2005.

[158] G. S. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, Feb 2005.

[159] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007.

[160] Andrews S. Fastqc: a quality control tool for high throughput sequence data., 2010.

[161] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[162] M. Hummel, S. Bonnin, E. Lowy, and G. Roma. Teqc: an r-package for quality control in target capture experiments. *Bioinformatics*, (27(9):1316-1317), 2011.

[163] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[164] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[165] Broad Institute. Picard tools, (Accessed: 2018/02/21; version 2.17.8).

[166] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

[167] DAVID J. STATES and WARREN GISH. QGB: Combined use of sequence similarity and codon bias for coding region identification. *Journal of Computational Biology*, 1(1):39–50, jan 1994.

[168] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, 4(8):1184–1191, 2009.

[169] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440, 2005.

[170] Guido van Rossum. Python software foundation. python language reference, version 2.7. available at http://www.python.org.

[171] Hadley Wickham, Romain Francois, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2017. R package version 0.7.4.

[172] Florian Hahne and Robert Ivanek. *Statistical Genomics: Methods and Protocols*, chapter Visualizing Genomic Data Using Gviz and Bioconductor, pages 335–351. Springer New York, New York, NY, 2016.

[173] E. H. Trager, R. Khanna, A. Marrs, L. Siden, K. E. Branham, A. Swaroop, and J. E. Richards. Madeline 2.0 pde: a new program for local and web-based pedigree drawing. *Bioinformatics*, 23(14):1854–1856, Jul 2007.

[174] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2009.

[175] Stephen Turner. *qqman: Q-Q and Manhattan Plots for GWAS Data*, 2017. R package version 0.1.4.

[176] Courtney Landers. PhD thesis, 2018.

[177] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics*, 54:1–1, 06 2016.

[178]

[179] X. Jiao, B. T. Sherman, d. a. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806, Jul 2012.

[180] NCBI remap team. Nbci remap tool.

[181] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4(7):1073–1081, 2009.

[182] Y. M. Badke, R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics*, 13:24, Jan 2012.

[183] D. Shigemizu, Y. Momozawa, T. Abe, T. Morizono, K. A. Boroevich, S. Takata, K. Ashikawa, M. Kubo, and T. Tsunoda. Performance comparison of four commercial human whole-exome capture platforms. *Sci Rep*, 5:12742, Aug 2015.

[184]

[185]

[186] K. S. Wang, Y. Liu, C. Xu, X. Liu, and X. Luo. Family-based association analysis of NAV2 gene with the risk and age at onset of Alzheimer's disease. *J. Neuroimmunol.*, 310:60–65, Sep 2017.

[187] E. M. McNeill, K. P. Roos, D. Moechars, and M. Clagett-Dame. Nav2 is necessary for cranial nerve development and blood pressure regulation. *Neural Dev*, 5:6, Feb 2010.

[188] E. M. McNeill, M. Klockner-Bormann, E. C. Roesler, L. E. Talton, D. Moechars, and M. Clagett-Dame. Nav2 hypomorphic mutant mice are ataxic and exhibit abnormalities in cerebellar development. *Dev. Biol.*, 353(2):331–343, May 2011.

[189] D. M. Dick, J. Meyers, F. Aliev, J. Nurnberger, J. Kramer, S. Kuperman, B. Porjesz, J. Tischfield, H. J. Edenberg, T. Foroud, M. Schuckit, A. Goate, V. Hesselbrock, and L. Bierut. Evidence for genes on chromosome 2 contributing to alcohol dependence with conduct disorder and suicide attempts. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 153B(6):1179–1188, Sep 2010.

[190] X. Wang, Y. Sekine, A. B. Byrne, W. B. Cafferty, M. Hammarlund, and S. M. Strittmatter. Inhibition of Poly-ADP-Ribosylation Fails to Increase Axonal Regeneration or Improve Functional Recovery after Adult Mammalian CNS Injury. *eNeuro*, 3(6), 2016.

[191] H. Inaba, A. Tsukagoshi, and S. Kida. PARP-1 activity is required for the reconsolidation and extinction of contextual fear memory. *Mol Brain*, 8(1):63, Oct 2015.

[192] C. Yu, B. S. Kim, and E. Kim. FAF1 mediates regulated necrosis through PARP1 activation upon oxidative stress leading to dopaminergic neurodegeneration. *Cell Death Differ.*, 23(11):1873–1885, 11 2016.

[193] S. Martire, L. Mosca, and M. d'Erme. PARP-1 involvement in neurodegeneration: A focus on Alzheimer's and Parkinson's diseases. *Mech. Ageing Dev.*, 146-148:53–64, Mar 2015.

[194] O. Engmann, T. Hortobagyi, A. J. Thompson, J. Guadagno, C. Troakes, S. Soriano, S. Al-Sarraj, Y. Kim, and K. P. Giese. Cyclin-dependent kinase 5 activator p25 is generated during memory formation and is reduced at an early stage in Alzheimer's disease. *Biol. Psychiatry*, 70(2):159–168, Jul 2011.

[195] M. Ide and D. A. Lewis. Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits. *Biol. Psychiatry*, 68(1):25–32, Jul 2010.

[196] P. T. Nelson, S. Estus, E. L. Abner, I. Parikh, M. Malik, J. H. Neltner, E. Ighodaro, W. X. Wang, B. R. Wilfred, L. S. Wang, W. A. Kukull, K. Nandakumar, M. L. Farman, W. W. Poon, M. M. Corrada, C. H. Kawas, D. H. Cribbs, D. A. Bennett, J. A. Schneider, E. B. Larson, P. K. Crane, O. Valladares, F. A. Schmitt, R. J. Kryscio, G. A. Jicha, C. D. Smith, S. W. Scheff, J. A. Sonnen, J. L. Haines, M. A. Pericak-Vance, R. Mayeux, L. A. Farrer, L. J. Van Eldik, C. Horbinski, R. C. Green, M. Gearing, L. W. Poon, P. L. Kramer, R. L. Woltjer, T. J. Montine, A. B. Partch, A. J. Rajic, K. Richmire, S. E. Monsell, G. D. Schellenberg, and D. W. Fardo.

ABCC9 gene polymorphism is associated with hippocampal sclerosis of aging pathology. *Acta Neuropathol.*, 127(6):825–843, 2014.

[197] Y. Majeed, S. Tumova, B. L. Green, V. A. Seymour, D. M. Woods, A. K. Agarwal, J. Naylor, S. Jiang, H. M. Picton, K. E. Porter, D. J. O'Regan, K. Muraki, C. W. Fishwick, and D. J. Beech. Pregnenolone sulphate-independent inhibition of TRPM3 channels by progesterone. *Cell Calcium*, 51(1):1–11, Jan 2012.

[198] A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G. B. Chen, V. Emilsson, S. F. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. van der Lee, C. deLeeuw, P. A. Lind, K. O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H. J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. B?nnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J. E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldorsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, J. J. Hottenga, P. L. de Jager, P. K. Joshi, A. Jugessur, M. A. Kaakinen, M. Kahonen, S. Kanoni, L. Keltigangas-Jarvinen, L. A. Kiemeney, I. Kolcic, S. Koskinen, A. T. Kraja, M. Kroh, Z. Kutalik, A. Latvala, L. J. Launer, M. P. Lebreton, D. F. Levinson, P. Lichtenstein, P. Lichtner, D. C. Liewald, A. Loukola, P. A. Madden, R. Magi, T. Maki-Opas, R. E. Marioni, P. Marques-Vidal, G. A. Meddens, G. McMahon, C. Meisinger, T. Meitinger, Y. Milaneschi, L. Milani, G. W. Montgomery, R. Myhre, C. P. Nelson, D. R. Nyholt, W. E. Ollier, A. Palotie, L. Paternoster, N. L. Pedersen, K. E. Petrovic, D. J. Porteous, K. Raikkonen, S. M. Ring, A. Robino, O. Rostapshova, I. Rudan, A. Rustichini, V. Salomaa, A. R. Sanders, A. P. Sarin, H. Schmidt, R. J. Scott, B. H. Smith, J. A. Smith, J. A. Staessen, E. Steinhagen-Thiessen, K. Strauch, A. Terracciano, M. D. Tobin, S. Ulivi, S. Vaccargiu, L. Quaye, F. J. van Rooij, C. Venturini, A. A. Vinkhuyzen, U. Volker, H. Volzke, J. M. Vonk, D. Vozzi, J. Waage, E. B. Ware, G. Willemsen, J. R. Attia, D. A. Bennett, K. Berger, L. Bertram, H. Bisgaard, D. I. Boomsma, I. B. Borecki, U. Bultmann, C. F. Chabris, F. Cucca, D. Cusi, I. J. Deary, G. V. Dedoussis, C. M. van Duijn, J. G. Eriksson, B. Franke, L. Franke, P. Gasparini, P. V. Gejman, C. Gieger, H. J. Grabe, J. Gratten, P. J. Groenen, V. Gudnason, P. van der Harst, C. Hayward, D. A. Hinds, W. Hoffmann, E. Hypponen,

W. G. Iacono, B. Jacobsson, M. R. Jarvelin, K. H. Jockel, J. Kaprio, S. L. Kardia, T. Lehtimaki, S. F. Lehrer, P. K. Magnusson, N. G. Martin, M. McGue, A. Metspalu, N. Pendleton, B. W. Penninx, M. Perola, N. Pirastu, M. Pirastu, O. Polasek, D. Posthuma, C. Power, M. A. Province, N. J. Samani, D. Schlessinger, R. Schmidt, T. I. S?rensen, T. D. Spector, K. Stefansson, U. Thorsteinsdottir, A. R. Thurik, N. J. Timpson, H. Tiemeier, J. Y. Tung, A. G. Uitterlinden, V. Vitart, P. Vollenweider, D. R. Weir, J. F. Wilson, A. F. Wright, D. C. Conley, R. F. Krueger, G. Davey Smith, A. Hofman, D. I. Laibson, S. E. Medland, M. N. Meyer, J. Yang, M. Johannesson, P. M. Visscher, T. Esko, P. D. Koellinger, D. Cesarini, D. J. Benjamin, B. Z. Alizadeh, R. A. de Boer, H. M. Boezen, M. Bruinenberg, L. Franke, P. van der Harst, H. L. Hillege, M. M. van der Klauw, G. Navis, J. Ormel, D. S. Postma, J. G. Rosmalen, J. P. Slaets, H. Snieder, R. P. Stolk, B. H. Wolffenbuttel, and C. Wijmenga. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539–542, 05 2016.

[199] M. Lam, J. W. Trampush, J. Yu, E. Knowles, G. Davies, D. C. Liewald, J. M. Starr, S. Djurovic, I. Melle, K. Sundet, A. Christoforou, I. Reinvang, P. DeRosse, A. J. Lundervold, V. M. Steen, T. Espeseth, K. Raikkonen, E. Widen, A. Palotie, J. G. Eriksson, I. Giegling, B. Konte, P. Roussos, S. Giakoumaki, K. E. Burdick, A. Payton, W. Ollier, O. Chiba-Falek, D. K. Attix, A. C. Need, E. T. Cirulli, A. N. Voineskos, N. C. Stefanis, D. Avramopoulos, A. Hatzimanolis, D. E. Arking, N. Smyrnis, R. M. Bilder, N. A. Freimer, T. D. Cannon, E. London, R. A. Poldrack, F. W. Sabb, E. Congdon, E. D. Conley, M. A. Scult, D. Dickinson, R. E. Straub, G. Donohoe, D. Morris, A. Corvin, M. Gill, A. R. Hariri, D. R. Weinberger, N. Pendleton, P. Bitsios, D. Rujescu, J. Lahti, S. Le Hellard, M. C. Keller, O. A. Andreassen, I. J. Deary, D. C. Glahn, A. K. Malhotra, and T. Lencz. Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets. *Cell Rep*, 21(9):2597–2613, Nov 2017.

[200] Z. Shang, H. Lv, M. Zhang, L. Duan, S. Wang, J. Li, G. Liu, Z. Ruijie, and Y. Jiang. Genome-wide haplotype association study identify TNFRSF1A, CASP7, LRP1B, CDH1 and TG genes associated with Alzheimer's disease in Caribbean Hispanic individuals. *Oncotarget*, 6(40):42504–42514, Dec 2015.

[201] S. E. Poduslo, R. Huang, and A. Spiro. A genome screen of successful aging without cognitive decline identifies LRP1B by haplotype analysis. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 153B(1):114–119, Jan 2010.

[202] C. Prats, B. Arias, J. Moya-Higueras, E. Pomarol-Clotet, M. Parellada, A. Gonzalez-Pinto, V. Peralta, M. I. Ibanez, M. Martin, L. Fananas, and M. Fatjo-Vilas. Evidence of an epistatic

effect between Dysbindin-1 and Neuritin-1 genes on the risk for schizophrenia spectrum disorders. *Eur. Psychiatry*, 40:60–64, 02 2017.

[203] D. Chandler, M. Dragovi?, M. Cooper, J. C. Badcock, B. H. Mullin, D. Faulkner, S. G. Wilson, J. Hallmayer, S. Howell, D. Rock, L. J. Palmer, L. Kalaydjieva, and A. Jablensky. Impact of Neuritin 1 (NRN1) polymorphisms on fluid intelligence in schizophrenia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 153B(2):428–437, Mar 2010.

[204] M. Fatjo-Vilas, C. Prats, E. Pomarol-Clotet, L. Lazaro, C. Moreno, I. Gonzalez-Ortega, S. Lera-Miguel, S. Miret, M. J. Munoz, I. Ibanez, S. Campanera, M. Giralt-Lopez, M. J. Cuesta, V. Peralta, G. Ortet, M. Parellada, A. Gonzalez-Pinto, P. J. McKenna, and L. Fananas. Involvement of NRN1 gene in schizophrenia-spectrum and bipolar disorders and its impact on age at onset and cognitive functioning. *World J. Biol. Psychiatry*, 17(2):129–139, 2016.

[205] J. D. Buxbaum, L. Georgieva, J. J. Young, C. Plescia, Y. Kajiwara, Y. Jiang, V. Moskvina, N. Norton, T. Peirce, H. Williams, N. J. Craddock, L. Carroll, G. Corfas, K. L. Davis, M. J. Owen, S. Harroch, T. Sakurai, and M. C. O'Donovan. Molecular dissection of NRG1-ERBB4 signaling implicates PTPRZ1 as a potential schizophrenia susceptibility gene. *Mol. Psychiatry*, 13(2):162–172, Feb 2008.

[206] B. Zhang, Y. H. Xu, S. G. Wei, H. B. Zhang, D. K. Fu, Z. F. Feng, F. L. Guan, Y. S. Zhu, and S. B. Li. Association study identifying a new susceptibility gene AUTS2 for schizophrenia. *Int J Mol Sci*, 15(11):19406–19416, Oct 2014.

[207] Y. Fan, W. Qiu, L. Wang, X. Gu, and Y. Yu. Exonic deletions of AUTS2 in Chinese patients with developmental delay and intellectual disability. *Am. J. Med. Genet. A*, 170A(2):515–522, Feb 2016.

[208] K. Hori, T. Nagai, W. Shan, A. Sakamoto, M. Abe, M. Yamazaki, K. Sakimura, K. Yamada, and M. Hoshino. Heterozygous Disruption of Autism susceptibility candidate 2 Causes Impaired Emotional Control and Cognitive Memory. *PLoS ONE*, 10(12):e0145979, 2015.

[209] F. Bedogni, R. D. Hodge, B. R. Nelson, E. A. Frederick, N. Shiba, R. A. Daza, and R. F. Hevner. Autism susceptibility candidate 2 (Auts2) encodes a nuclear protein expressed in developing brain regions implicated in autism neuropathology. *Gene Expr. Patterns*, 10(1):9–15, Jan 2010.

[210] L. Li, H. Sun, J. Ding, C. Niu, M. Su, L. Zhang, Y. Li, C. Wang, N. Gamper, X. Du, and H. Zhang. Selective targeting of M-type potassium Kv 7.4 channels demonstrates their key role in the regulation of dopaminergic neuronal excitability and depression-like behaviour. *Br. J. Pharmacol.*, 174(23):4277–4294, Dec 2017.

[211] A. J. Griswold, D. Ma, S. J. Sacharow, J. L. Robinson, J. M. Jaworski, H. H. Wright, R. K. Abramson, H. Lybaek, N. ?yen, M. L. Cuccaro, J. R. Gilbert, and M. A. Pericak-Vance. A de novo 1.5 Mb microdeletion on chromosome 14q23.2-23.3 in a patient with autism and spherocytosis. *Autism Res*, 4(3):221–227, Jun 2011.

[212] W. S. Kang, J. K. Park, S. K. Kim, H. J. Park, S. M. Lee, J. Y. Song, J. H. Chung, and J. W. Kim. Genetic variants of GRIA1 are associated with susceptibility to schizophrenia in Korean population. *Mol. Biol. Rep.*, 39(12):10697–10703, Dec 2012.

[213] B. Kerner, A. J. Jasinska, J. DeYoung, M. Almonte, O. W. Choi, and N. B. Freimer. Polymorphisms in the GRIA1 gene region in psychotic bipolar disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 150B(1):24–32, Jan 2009.

[214] Illumina. Calculating percent passing filter for patterned and non-patterned flow cells: A comparison of methods for calculating percent passing filter on illumina flow cells. Illumina White paper, 2015.

[215] J. Shang, F. Zhu, W. Vongsangnak, Y. Tang, W. Zhang, and B. Shen. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res Int*, 2014:309650, 2014.

[216] A. Rhoads and K. F. Au. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*, 13(5):278–289, Oct 2015.

[217] L. B. Schook, T. V. Collares, K. A. Darfour-Oduro, A. K. De, L. A. Rund, K. M. Schachtschneider, and F. K. Seixas. Unraveling the swine genome: implications for human health. *Annu Rev Anim Biosci*, 3:219–244, 2015.

[218] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, Nov 2008.

[219] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Benjamin Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 2017.

[220] M. Kannan, E. Bayam, C. Wagner, B. Rinaldi, P. F. Kretz, P. Tilly, M. Roos, L. McGillewie, S. Bar, S. Minocha, C. Chevalier, C. Po, J. Chelly, J. L. Mandel, R. Borgatti, A. Piton, C. Kinnear, B. Loos, D. J. Adams, Y. Herault, S. C. Collins, S. Friant, J. D. Godin, B. Yalcin, V. E. Vancollie, L. F. E. Anthony, S. A. Maguire, D. Lafont, S. A. Pearson, A. S. Gates, M. Sanderson,

C. Shannon, M. T. Sumowski, R. S. B. McLaren-Jones, A. Swiatkowska, C. M. Isherwood, E. L. Cambridge, H. M. Wilson, S. S. Caetano, A. K. B. Maguire, A. Galli, A. O. Speak, J. Dench, E. Tuck, J. Estabel, A. Green, C. Tudor, E. Siragher, M. Dabrowska, C. I. Mazzeo, Y. Hooks, F. Kussy, M. Griffiths, D. Gannon, B. Doe, K. Boroviak, H. Wardle-Jones, N. Griggs, J. Bottomley, E. Ryder, D. Gleeson, J. K. White, R. Ramirez-Solis, and C. J. Lelliott. WD40-repeat 47, a microtubule-associated protein, is essential for brain development and autophagy. *Proc. Natl. Acad. Sci. U.S.A.*, 114(44):E9308–E9317, 10 2017.

[221] F. S. Goes, J. McGrath, D. Avramopoulos, P. Wolyniec, M. Pirooznia, I. Ruczinski, G. Nestadt, E. E. Kenny, V. Vacic, I. Peters, T. Lencz, A. Darvasi, J. G. Mulle, S. T. Warren, and A. E. Pulver. Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 168(8):649–659, Dec 2015.

[222] B. Smolarz, D. Skalski, A. Rysz, A. Marchel, H. Romanowicz, and M. Makowska. Polymorphism of the multidrug resistance 1 gene MDR1 G2677T/A (rs2032582) and the risk of drug-resistant epilepsy in the Polish adult population. *Acta Neurol Belg*, 117(4):849–855, Dec 2017.

[223] R. J. Lv, X. Q. Shao, T. Cui, and Q. Wang. Significance of MDR1 gene C3435T polymorphism in predicting childhood refractory epilepsy. *Epilepsy Res.*, 132:21–28, 05 2017.

[224] X. Zhong, M. Y. Liu, X. H. Sun, and M. J. Wei. Association between ABCB1 polymorphisms and haplotypes and Alzheimer's disease: a meta-analysis. *Sci Rep*, 6:32708, 09 2016.

[225] J. A. Hashmi, K. M. Al-Harbi, K. Ramzan, A. M. Albalawi, A. Mehmood, M. I. Samman, and S. Basit. A novel splice-site mutation in the ASPM gene underlies autosomal recessive primary microcephaly. *Ann Saudi Med*, 36(6):391–396, 2016.

[226] M. S. Abdel-Hamid, M. F. Ismail, H. A. Darwish, L. K. Effat, M. S. Zaki, and G. M. Abdel-Salam. Molecular and phenotypic spectrum of ASPM-related primary microcephaly: Identification of eight novel mutations. *Am. J. Med. Genet. A*, 170(8):2133–2140, 08 2016.

[227] C. M. Stein, B. Truitt, F. Deng, A. A. Ciesla, F. Qiu, P. Joseph, R. Raghavendra, J. Fondran, R. P. Igo, J. Tag, L. Freebairn, H. G. Taylor, B. A. Lewis, and S. K. Iyengar. Association between AVPR1A, DRD2, and ASPM and endophenotypes of communication disorders. *Psychiatr. Genet.*, 24(5):191–200, Oct 2014.

[228] M. R. Cohen, W. M. Johnson, J. M. Pilat, J. Kiselar, A. DeFrancesco-Lisowitz, R. E. Zigmond, and V. Y. Moiseenkova-Bell. Nerve Growth Factor Regulates Transient Receptor Potential Vanilloid 2 via Extracellular Signal-Regulated Kinase Signaling To Enhance Neurite Outgrowth in Developing Neurons. *Mol. Cell. Biol.*, 35(24):4238–4252, Dec 2015.

[229] Jingyu Liu, Jennifer Ciarochi, Vince D. Calhoun, Jane S. Paulsen, H. Jeremy Bockholt, Hans J. Johnson, Jeffrey D. Long, Dongdong Lin, Flor A. Espinoza, Maria B. Misiura, Arvind Caprihan, and Jessica A. Turner and. Genetics modulate gray matter variation beyond disease burden in prodromal huntington's disease. *Frontiers in Neurology*, 9, mar 2018.

[230] M. J. Geiger, K. Domschke, G. A. Homola, S. M. Schulz, J. Nowak, A. Akhrif, P. Pauli, J. Deckert, and S. Neufang. ADORA2A genotype modulates interoceptive and exteroceptive processing in a fronto-insular network. *Eur Neuropsychopharmacol*, 26(8):1274–1285, 08 2016.

[231] E. Horgusluoglu-Moloch, K. Nho, S. L. Risacher, S. Kim, T. Foroud, L. M. Shaw, J. Q. Trojanowski, P. S. Aisen, R. C. Petersen, C. R. Jack, S. Lovestone, A. Simmons, M. W. Weiner, and A. J. Saykin. Targeted neurogenesis pathway-based gene analysis identifies ADORA2A associated with hippocampal volume in mild cognitive impairment and Alzheimer's disease. *Neurobiol. Aging*, 60:92–103, Dec 2017.

[232] J. Stockwell, E. Jakova, and F. S. Cayabyab. Adenosine A1 and A2A Receptors in the Brain: Current Research and Their Role in Neurodegeneration. *Molecules*, 22(4), Apr 2017.

[233] M. Aspe-Sanchez, M. Moreno, M. I. Rivera, A. Rossi, and J. Ewer. Oxytocin and Vasopressin Receptor Gene Polymorphisms: Role in Social and Psychiatric Traits. *Front Neurosci*, 9:510, 2015.

[234] D. van West, J. Del-Favero, Y. Aulchenko, P. Oswald, D. Souery, T. Forsgren, S. Sluijs, S. Bel-Kacem, R. Adolfsson, J. Mendlewicz, C. Van Duijn, D. Deboutte, C. Van Broeckhoven, and S. Claes. A major SNP haplotype of the arginine vasopressin 1B receptor protects against recurrent major depression. *Mol. Psychiatry*, 9(3):287–292, Mar 2004.

[235] D. Luppino, C. Moul, D. J. Hawes, J. Brennan, and M. R. Dadds. Association between a polymorphism of the vasopressin 1B receptor gene and aggression in children. *Psychiatr. Genet.*, 24(5):185–190, Oct 2014.

[236] N. J. Grundwald, D. P. Benitez, and P. J. Brunton. Sex-Dependent Effects of Prenatal Stress on Social Memory in Rats: A Role for Differential Expression of Central Vasopressin-1a Receptors. *J. Neuroendocrinol.*, 28(4), 04 2016.

[237] Y. Jia, T. J. Jucius, S. A. Cook, and S. L. Ackerman. Loss of Clcc1 results in ER stress, misfolded protein accumulation, and neurodegeneration. *J. Neurosci.*, 35(7):3001–3009, Feb 2015.

[238] F. Segovia-Miranda, F. Serrano, A. Dyrda, E. Ampuero, C. Retamal, M. Bravo-Zehnder, J. Parodi, P. Zamorano, D. Valenzuela, L. Massardo, B. van Zundert, N. C. Inestrosa, and A. Gonzalez. Pathogenicity of lupus anti-ribosomal P antibodies: role of cross-reacting neuronal surface

P antigen in glutamatergic transmission and plasticity in a mouse model. *Arthritis Rheumatol*, 67(6):1598–1610, Jun 2015.

[239] H. Morino, R. Miyamoto, S. Ohnishi, H. Maruyama, and H. Kawakami. Exome sequencing reveals a novel TTC19 mutation in an autosomal recessive spinocerebellar ataxia patient. *BMC Neurol*, 14:5, Jan 2014.

[240] D. Ghezzi, P. Arzuffi, M. Zordan, C. Da Re, C. Lamperti, C. Benna, P. D'Adamo, D. Diodato, R. Costa, C. Mariotti, G. Uziel, C. Smiderle, and M. Zeviani. Mutations in TTC19 cause mitochondrial complex III deficiency and neurological impairment in humans and flies. *Nat. Genet.*, 43(3):259–263, Mar 2011.

[241] R. H. Perlis, J. Huang, S. Purcell, M. Fava, A. J. Rush, P. F. Sullivan, S. P. Hamilton, F. J. McMahon, T. G. Schulze, T. Schulze, J. B. Potash, P. P. Zandi, V. L. Willour, B. W. Penninx, D. I. Boomsma, N. Vogelzangs, C. M. Middeldorp, M. Rietschel, M. Nothen, S. Cichon, H. Gurling, N. Bass, A. McQuillin, M. Hamshere, N. Craddock, P. Sklar, and J. W. Smoller. Genome-wide association study of suicide attempts in mood disorder patients. *Am J Psychiatry*, 167(12):1499–1507, Dec 2010.

[242] Q. S. Li, C. Tian, G. R. Seabrook, W. C. Drevets, and V. A. Narayan. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. *Transl Psychiatry*, 6(9):e889, 09 2016.

[243] Tsutomu Mori, Yuanyuan Li, Hiroaki Hata, Kazuo Ono, and Hideo Kochi. NIRF, a novel RING finger protein, is involved in cell-cycle regulation. *Biochemical and Biophysical Research Communications*, 296(3):530–536, aug 2002.

[244] John Neidhardt, Susanne Fehr, Michael Kutsche, Jürgen Löhler, and Melitta Schachner. Tenascin-n: characterization of a novel member of the tenascin family that mediates neurite repulsion from hippocampal explants. *Molecular and Cellular Neuroscience*, 23(2):193–209, jun 2003.

[245] P. F. Sullivan, A. Agrawal, C. M. Bulik, O. A. Andreassen, A. D. B?rglum, G. Breen, S. Cichon, H. J. Edenberg, S. V. Faraone, J. Gelernter, C. A. Mathews, C. M. Nievergelt, J. W. Smoller, and M. C. O'Donovan. Psychiatric Genomics: An Update and an Agenda. *Am J Psychiatry*, 175(1):15–27, Jan 2018.

[246] R. A. Power, K. E. Tansey, H. N. Buttensch?n, S. Cohen-Woods, T. Bigdeli, L. S. Hall, Z. Kutalik, S. H. Lee, S. Ripke, S. Steinberg, A. Teumer, A. Viktorin, N. R. Wray, V. Arolt, B. T. Baune, D. I. Boomsma, A. D. B?rglum, E. M. Byrne, E. Castelao, N. Craddock, I. W. Craig, U. Dannlowski, I. J. Deary, F. Degenhardt, A. J. Forstner, S. D. Gordon, H. J. Grabe, J. Grove,

S. P. Hamilton, C. Hayward, A. C. Heath, L. J. Hocking, G. Homuth, J. J. Hottenga, S. Kloiber, J. Krogh, M. Landen, M. Lang, D. F. Levinson, P. Lichtenstein, S. Lucae, D. J. MacIntyre, P. Madden, P. K. E. Magnusson, N. G. Martin, A. M. McIntosh, C. M. Middeldorp, Y. Milaneschi, G. W. Montgomery, O. Mors, B. Muller-Myhsok, D. R. Nyholt, H. Oskarsson, M. J. Owen, S. Padmanabhan, B. W. J. H. Penninx, M. L. Pergadia, D. J. Porteous, J. B. Potash, M. Preisig, M. Rivera, J. Shi, S. I. Shyn, E. Sigurdsson, J. H. Smit, B. H. Smith, H. Stefansson, K. Stefansson, J. Strohmaier, P. F. Sullivan, P. Thomson, T. E. Thorgeirsson, S. Van der Auwera, M. M. Weissman, G. Breen, and C. M. Lewis. Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. *Biol. Psychiatry*, 81(4):325–335, 02 2017.

[247] J. A. Lee, A. Damianov, C. H. Lin, M. Fontes, N. N. Parikshak, E. S. Anderson, D. H. Geschwind, D. L. Black, and K. C. Martin. Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. *Neuron*, 89(1):113–128, Jan 2016.

[248] M. N. Davies, S. Verdi, A. Burri, M. Trzaskowski, M. Lee, J. M. Hettema, R. Jansen, D. I. Boomsma, and T. D. Spector. Generalised Anxiety Disorder–A Twin Study of Genetic Architecture, Genome-Wide Association and Differential Gene Expression. *PLoS ONE*, 10(8):e0134865, 2015.

[249] S. J. Mosca, L. M. Langevin, D. Dewey, A. M. Innes, A. C. Lionel, C. C. Marshall, S. W. Scherer, J. S. Parboosingh, and F. P. Bernier. Copy-number variations are enriched for neurodevelopmental genes in children with developmental coordination disorder. *J. Med. Genet.*, 53(12):812–819, 12 2016.

[250] J. A. Lee, Z. Z. Tang, and D. L. Black. An inducible change in Fox-1/A2BP1 splicing modulates the alternative splicing of downstream neuronal target exons. *Genes Dev.*, 23(19):2284–2293, Oct 2009.

[251] E. E. Custer, T. K. Knott, A. E. Cuadra, S. Ortiz-Miranda, and J. R. Lemos. P2X purinergic receptor knockout mice reveal endogenous ATP modulation of both vasopressin and oxytocin release from the intact neurohypophysis. *J. Neuroendocrinol.*, 24(4):674–680, Apr 2012.

[252] Mayada Tassabehji, Andrew P. Read, Valerie E. Newton, Rodney Harris, Rudi Balling, Peter Gruss, and Tom Strachan. Waardenburg's syndrome patients have mutations in the human homologue of the pax-3 paired box gene. *Nature*, 355(6361):635–636, February 1992.

[253] M. M. Davis, P. Olausson, P. Greengard, J. R. Taylor, and A. C. Nairn. Regulator of calmodulin signaling knockout mice display anxiety-like behavior and motivational deficits. *Eur. J. Neurosci.*, 35(2):300–308, Jan 2012.

[254] G. Marangi, D. Orteschi, V. Milano, G. Mancano, and M. Zollino. Interstitial deletion of 3p22.3p22.2 encompassing ARPP21 and CLASP2 is a potential pathogenic factor for a syndromic form of intellectual disability: a co-morbidity model with additional copy number variations in a large family. *Am. J. Med. Genet. A*, 161A(11):2890–2893, Nov 2013.

[255] T. W. Fitzgerald, S. S. Gerety, W. D. Jones, M. van Kogelenberg, D. A. King, J. McRae, K. I. Morley, V. Parthiban, S. Al-Turki, K. Ambridge, D. M. Barrett, T. Bayzetinova, S. Clayton, E. L. Coomber, S. Gribble, P. Jones, N. Krishnappa, L. E. Mason, A. Middleton, R. Miller, E. Prigmore, D. Rajan, A. Sifrim, A. R. Tivey, M. Ahmed, N. Akawi, R. Andrews, U. Anjum, H. Archer, R. Armstrong, M. Balasubramanian, R. Banerjee, D. Baralle, P. Batstone, D. Baty, C. Bennett, J. Berg, B. Bernhard, A. P. Bevan, E. Blair, M. Blyth, D. Bohanna, L. Bourdon, D. Bourn, A. Brady, E. Bragin, C. Brewer, L. Brueton, K. Brunstrom, S. J. Bumpstead, D. J. Bunyan, J. Burn, J. Burton, N. Canham, B. Castle, K. Chandler, S. Clasper, J. Clayton-Smith, T. Cole, A. Collins, M. N. Collinson, F. Connell, N. Cooper, H. Cox, L. Cresswell, G. Cross, Y. Crow, M. D'Alessandro, T. Dabir, R. Davidson, S. Davies, J. Dean, C. Deshpande, G. Devlin, A. Dixit, A. Dominiczak, C. Donnelly, D. Donnelly, A. Douglas, A. Duncan, J. Eason, S. Edkins, S. Ellard, P. Ellis, F. Elmslie, K. Evans, S. Everest, T. Fendick, R. Fisher, F. Flinter, N. Foulds, A. Fryer, B. Fu, C. Gardiner, L. Gaunt, N. Ghali, R. Gibbons, S. L. Gomes Pereira, J. Goodship, D. Goudie, E. Gray, P. Greene, L. Greenhalgh, L. Harrison, R. Hawkins, S. Hellens, A. Henderson, E. Hobson, S. Holden, S. Holder, G. Hollingsworth, T. Homfray, M. Humphreys, J. Hurst, S. Ingram, M. Irving, J. Jarvis, L. Jenkins, D. Johnson, D. Jones, E. Jones, D. Josifova, S. Joss, B. Kaemba, S. Kazembe, B. Kerr, U. Kini, E. Kinning, G. Kirby, C. Kirk, E. Kivuva, A. Kraus, D. Kumar, K. Lachlan, W. Lam, A. Lampe, C. Langman, M. Lees, D. Lim, G. Lowther, S. A. Lynch, A. Magee, E. Maher, S. Mansour, K. Marks, K. Martin, U. Maye, E. McCann, V. McConnell, M. McEntagart, R. McGowan, K. McKay, S. McKee, D. J. McMullan, S. McNerlan, S. Mehta, K. Metcalfe, E. Miles, S. Mohammed, T. Montgomery, D. Moore, S. Morgan, A. Morris, J. Morton, H. Mugalaasi, V. Murday, L. Nevitt, R. Newbury-Ecob, A. Norman, R. O'Shea, C. Ogilvie, S. Park, M. J. Parker, C. Patel, J. Paterson, S. Payne, J. Phipps, D. T. Pilz, D. Porteous, N. Pratt, K. Prescott, S. Price, A. Pridham, A. Procter, H. Purnell, N. Ragge, J. Rankin, L. Raymond, D. Rice, L. Robert, E. Roberts, G. Roberts, J. Roberts, P. Roberts, A. Ross, E. Rosser, A. Saggar, S. Samant, R. Sandford, A. Sarkar, S. Schweiger, C. Scott, R. Scott, A. Selby, A. Seller, C. Sequeira, N. Shannon, S. Sharif, C. Shaw-Smith, E. Shearing, D. Shears, I. Simonic, D. Simpkin, R. Singzon, Z. Skitt, A. Smith, B. Smith, K. Smith, S. Smithson, L. Sneddon, M. Splitt, M. Squires, F. Stewart, H. Stewart, M. Suri, V. Sutton, G. J. Swaminathan, E. Sweeney, K. Tatton-Brown, C. Taylor, R. Taylor, M. Tein, I. K. Temple, J. Thomson, J. Tolmie, A. Torokwa, B. Treacy, C. Turner, P. Turnpenny, C. Tysoe, A. Vandersteen, P. Vas-

udevan, J. Vogt, E. Wakeling, D. Walker, J. Waters, A. Weber, D. Wellesley, M. Whiteford, S. Widaa, S. Wilcox, D. Williams, N. Williams, G. Woods, C. Wragg, M. Wright, F. Yang, M. Yau, N. P. Carter, M. Parker, H. V. Firth, D. R. FitzPatrick, C. F. Wright, J. C. Barrett, and M. E. Hurles. Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228, Mar 2015.

[256] C. Herold, B. V. Hooli, K. Mullin, T. Liu, J. T. Roehr, M. Mattheisen, A. R. Parrado, L. Bertram, C. Lange, and R. E. Tanzi. Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. *Mol. Psychiatry*, 21(11):1608–1612, 11 2016.

[257] M. A. Corbett, S. T. Bellows, M. Li, R. Carroll, S. Micallef, G. L. Carvill, C. T. Myers, K. B. Howell, S. Maljevic, H. Lerche, E. V. Gazina, H. C. Mefford, M. Bahlo, S. F. Berkovic, S. Petrou, I. E. Scheffer, and J. Gecz. Dominant KCNA2 mutation causes episodic ataxia and pharmacoresponsive epilepsy. *Neurology*, 87(19):1975–1984, Nov 2016.

[258] K. L. Helbig, U. B. Hedrich, D. N. Shinde, I. Krey, A. C. Teichmann, J. Hentschel, J. Schubert, A. C. Chamberlin, R. Huether, H. M. Lu, W. A. Alcaraz, S. Tang, C. Jungbluth, S. L. Dugan, L. Vainionpaa, K. N. Karle, M. Synofzik, L. Schols, R. Schule, A. E. Lehesjoki, I. Helbig, H. Lerche, and J. R. Lemke. A recurrent mutation in KCNA2 as a novel cause of hereditary spastic paraplegia and ataxia. *Ann. Neurol.*, 80(4), 10 2016.

[259] J. Li, R. P. Hart, E. M. Mallimo, M. R. Swerdel, A. W. Kusnecov, and K. Herrup. EZH2-mediated H3K27 trimethylation mediates neurodegeneration in ataxia-telangiectasia. *Nat. Neurosci.*, 16(12):1745–1753, Dec 2013.

[260] K. J. Billingsley, M. Manca, O. Gianfrancesco, D. A. Collier, H. Sharp, V. J. Bubb, and J. P. Quinn. Regulatory characterisation of the schizophrenia-associated CACNA1C proximal promoter and the potential role for the transcription factor EZH2 in schizophrenia aetiology. *Schizophr. Res.*, Feb 2018.

[261] J. C. Lambert, B. Grenier-Boley, D. Harold, D. Zelenika, V. Chouraki, Y. Kamatani, K. Sleegers, M. A. Ikram, M. Hiltunen, C. Reitz, I. Mateo, T. Feulner, M. Bullido, D. Galimberti, L. Concari, V. Alvarez, R. Sims, A. Gerrish, J. Chapman, C. Deniz-Naranjo, V. Solfrizzi, S. Sorbi, B. Arosio, G. Spalletta, G. Siciliano, J. Epelbaum, D. Hannequin, J. F. Dartigues, C. Tzourio, C. Berr, E. M. Schrijvers, R. Rogers, G. Tosto, F. Pasquier, K. Bettens, C. Van Cauwenberghe, L. Fratiglioni, C. Graff, M. Delepine, R. Ferri, C. A. Reynolds, L. Lannfelt, M. Ingelsson, J. A. Prince, C. Chillotti, A. Pilotto, D. Seripa, A. Boland, M. Mancuso, P. Bossu, G. Annoni, B. Nacmias, P. Bosco, F. Panza, F. Sanchez-Garcia, M. Del Zompo, E. Coto, M. Owen, M. O'Donovan,

F. Valdivieso, P. Caffarra, P. Caffara, E. Scarpini, O. Combarros, L. Buee, D. Campion, H. Soininen, M. Breteler, M. Riemenschneider, C. Van Broeckhoven, A. Alperovitch, M. Lathrop, D. A. Tregouet, J. Williams, and P. Amouyel. Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease. *Mol. Psychiatry*, 18(4):461–470, Apr 2013.

[262] L. P. Sutton, C. Orlandi, C. Song, W. C. Oh, B. S. Muntean, K. Xie, A. Filippini, X. Xie, R. Satterfield, J. D. W. Yaeger, K. J. Renner, S. M. Young, B. Xu, H. Kwon, and K. A. Martemyanov. Orphan receptor GPR158 controls stress-induced depression. *Elife*, 7, Feb 2018.

[263] F J Hosking, D Feldman, R Bruchim, B Olver, A Lloyd, J Vijayakrishnan, P Flint-Richter, P Broderick, R S Houlston, and S Sadetzki. Search for inherited susceptibility to radiation-associated meningioma by genomewide SNP linkage disequilibrium mapping. *British Journal of Cancer*, 104(6):1049–1054, mar 2011.

[264] Janice Estus. Combining genetic association study designs: a GWAS case study. *Frontiers in Genetics*, 4, 2013.

[265] Benjamin Georgi, David Craig, Rachel L. Kember, Wencheng Liu, Ingrid Lindquist, Sara Nasser, Christopher Brown, Janice A. Egeland, Steven M. Paul, and Maja Bućan. Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genetics*, 10(3):e1004229, mar 2014.