Genetics of Ivermectin Response and Linkage Disequilibrium in *Onchocerca volvulus*

Submitted by

Olayemi Mary Awobifa

Bachelor of Science (Applied Zoology), 2009 Olabisi Onabanjo University

Master of Science (Entomology), 2014, University of Ibadan

Master of Information Technology, 2014 Ladoke Akintola University of Technology

A thesis submitted in total fulfilment
of the requirements for the degree of

Doctor of Philosophy

School of Life Sciences
College of Science, Health and Engineering

La Trobe University
Victoria, Australia

November 2020

Table of Contents

Lists of Figures.

Lists of Tables.

Appendices

Abbreviations

APOC	-	African Programme for Onchocerciasis Control
ABZ	-	Albendazole
°C	-	degrees Celsius
Bp	-	Base pairs
BZ	-	Benzimidazoles
ESP	-	Exome Sequencing Project
ESPEN	-	Expanded Special Project for the Elimination of Neglected Tropical Diseases in Africa
F1	-	First filial generation
F2	-	Second Filial Generation
GR	-	Good responder
GWAS	-	Genome-wide Association Study
IDT	-	Integrated DNA technologies
Kb	-	Kilobase
LD	-	Linkage Disequilibrium
LOE	-	Loss of Efficacy
Mb	-	Mega base
MDA	-	Mass Drug Administration
ML	-	Macrocyclic Lactone
NCBI	-	National Centre for Biotechnology Information
N_e	-	effective population size
NGS	-	Next-Generation Sequencing
NTDs	-	Neglected Tropical Diseases
OCP	-	Onchocerciasis Control Programme
OEPA	-	Onchocerciasis Elimination Programme for the Americas
Pool-seq	-	Pooled next generation sequencing
QTLs	-	Quantitative trait loci
SD	-	standard deviation
SNPs	-	Single Nucleotide Polymorphisms
SOR	-	sub-optimal response
WHO	-	World Health Organisation
µg	-	microgram
µl	-	micro litre
µM	-	micro molar

Abstract

Onchocerca volvulus, the cause of river blindness, shows variation in response to the drug used to treat it (ivermectin). Genome Wide Association Studies (GWAS) revealed association between sub-optimal response (SOR) and genotype in adult female worms and suggested that SOR is a polygenic quantitative trait. However, previous genome sequencing either used pooled DNA or a limited number of single worms, due to a small number of worms and because of DNA quality. The primary aim of this thesis was to improve diagnostic capability of SOR in *O. volvulus* and aid elimination goals. In Chapter 2, I took an amplicon approach to validate candidates and improve sample size. I reported the findings from a pilot study that aimed to test whether 26 non-synonymous SNPs that fell within an ~7kb region from 14 quantitative trait loci (QTLs) defined in the GWAS were associated with ivermectin response. None of those chosen SNPs were in strong association with SOR and were therefore not predictive of ivermectin response by *O. volvulus* and unlikely to play a causative role in the SOR phenotype. However, the evolution study showed variation in selection across the loci studied. In Chapter 3, I explored an alternative approach to laborious resequencing of SNPs by exploring the localised and broad patterns of linkage disequilibrium (LD) and haploblock structure in the *O. volvulus* genome. The study gave a detailed insight into the SNP density required for designing a SNP array for future use in developing diagnostic tools. The genome-wide pattern of LD also confirmed that soft selection was driving SOR in *O. volvulus* and it left a weak LD signature in the genome which was characterised by clusters of small fragmented haploblocks of low to moderately elevated LD that correlate with peaks of $F_{ST.}$ In Chapter 4, I explored the feasibility and success of genomic imputation (a novel tool in helminth parasites genomics) in improving the density of SNPs available for GWAS. Evidence shows that genomic imputation could improve the probabilities of association for GWAS and could be employed in helminth's genomics at large. However, a larger reference panel that takes population structure into account should be considered. The findings in this thesis have, therefore, provided new insights into the essential requirements needed to finally eliminate onchocerciasis from Africa.

Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma.

No other person's work has been used without due acknowledgement in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

In chapter three and four of this thesis, Illumina HiSeq paired end, 100-bp sequence data for *O. volvulus* worms from Ghana, Mali, and Cote d'Ivoire; reduced representation and whole genome Illumina NextSeq sequenced reads from Cameroon were obtained from databases available from the Nematode Functional Genomics Laboratory (Grant Lab) at La Trobe University.

This work was supported by a La Trobe University Postgraduate Research Scholarship (LTUPRS) and a La Trobe University Full Fee Research Scholarship (LTUFFRS).

This research was funded by World Health Organization Special Programme for Research and Training in Tropical Diseases (WHO/TDR).

Olayemi Mary Awobifa 03 November 2020

Acknowledgements

First and foremost, all glory and honour to God Almighty for His abundant mercy bestowed on me throughout my PhD study and for His grace to complete this research successfully.

I would like to thank the following people for their valuable support in making this thesis a reality. My deep and sincere gratitude goes to my principal supervisor, Assoc. Prof. Warwick Grant for his supervision of my PhD project, enlightening discussions, and constructive feedback. I am also highly appreciative of my co-supervisor, Dr. Shannon Hedtke for her constant encouragement, guidance, and prompt advice. I would like to thank my research panel chair, Dr. Hans Daetwyler for his useful advice and support during my PhD study.

I would like to thank all the past and present Grant Lab members. Special thanks to Ernest and Sheila for helping me get started in the lab, Katie for helping me out when bioinformatics gets tough, Gowtam and Himal for their words of encouragement.

To my family, a huge thank you to my husband and son, Kay and Jason, for their love, understanding, prayers, continual support and inspiration to be diligent throughout the period of this research. Also, a big thank you to my parents and siblings, for their love, valuable prayers, care, and sacrifices.

A special thanks all the members of Deeper Christian Life Ministry, especially my pastors for their spiritual and mental support throughout my study.

Chapter One

General Introduction.

1.1.*Onchocerca volvulus* – The parasite, disease, control, and elimination.

Onchocerca volvulus is a parasitic nematode which causes the disease onchocerciasis (known commonly as river blindness). The disease is called river blindness because the blackfly that transmits the infection lives and breeds near fast-flowing streams and rivers, mostly near remote rural agricultural areas where at-risk people live and work (Centers for Disease Control, 2019), and because the most severe pathology caused by the disease is irreversible blindness (Little et al., 2004).

Onchocerciasis is among the Neglected Tropical Diseases (NTDs) with its main burden in 31 countries in sub-Saharan Africa and some parts of South America and in Yemen in the Middle East. Onchocerciasis is a serious public health problem and the second leading cause of infectious blindness in Africa (World Health Organization, 2019). At least, 20.9 million people are infected worldwide, of which 14.6 million have skin disease and 1.15 million have vision loss (Centers for Disease Control, 2019).

Endemicity levels of onchocerciasis comprises of three categories: *first, hypoendemic* – areas or foci where nodule prevalence is found in <20% of adults and corresponds to skin microfilarial prevalence of <30–35%, *second, mesoendemic* - areas or foci where nodule prevalence is between 20-40% of adults and skin microfilariae prevalence is between 30–35% and 60%, and *third, hyperendemic* - areas or foci where nodule prevalence is >40% of adults and skin microfilariae prevalence ≥60%. A microfilarial prevalence ≥80% has also been used to indicate *holoendemicity* (Noma et al., 2002, UNICEF and UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases, 1996).

Figure 1.1 shows the life cycle of the parasite - *O. volvulus* - as it migrates through its hosts at different filarial stages (Centers for Disease Control, 2019). The third-stage filarial larvae are transmitted by repeated bites of infected *Simulium* spp. blackflies into the human host (Figure 1.1) (Centers for Disease Control, 2019). In the human host, the third-stage filarial larvae develop into adult worms, which commonly reside in nodules under the skin and continue reproducing for approximately 15 years (Plaiser et al. 1991).

This is the maximum reproductive lifespan of the adult worm which is characterized by the age at which 95% of the adult worms have ceased reproduction and it is an important determinant of the period during which vector control must be continued to minimize the risk of recrudescence of onchocerciasis (Plaiser et al. 1991). The fertilised adult female worm's fecundity starts after a pre-mature period of one year on average (Duke, 1980). Following this period, microfilariae production only takes place when worms mate (Schulz-Key and Karam, 1986) and the microfilariae output decreases with age after a few years of productivity (Karam et al. 1987).

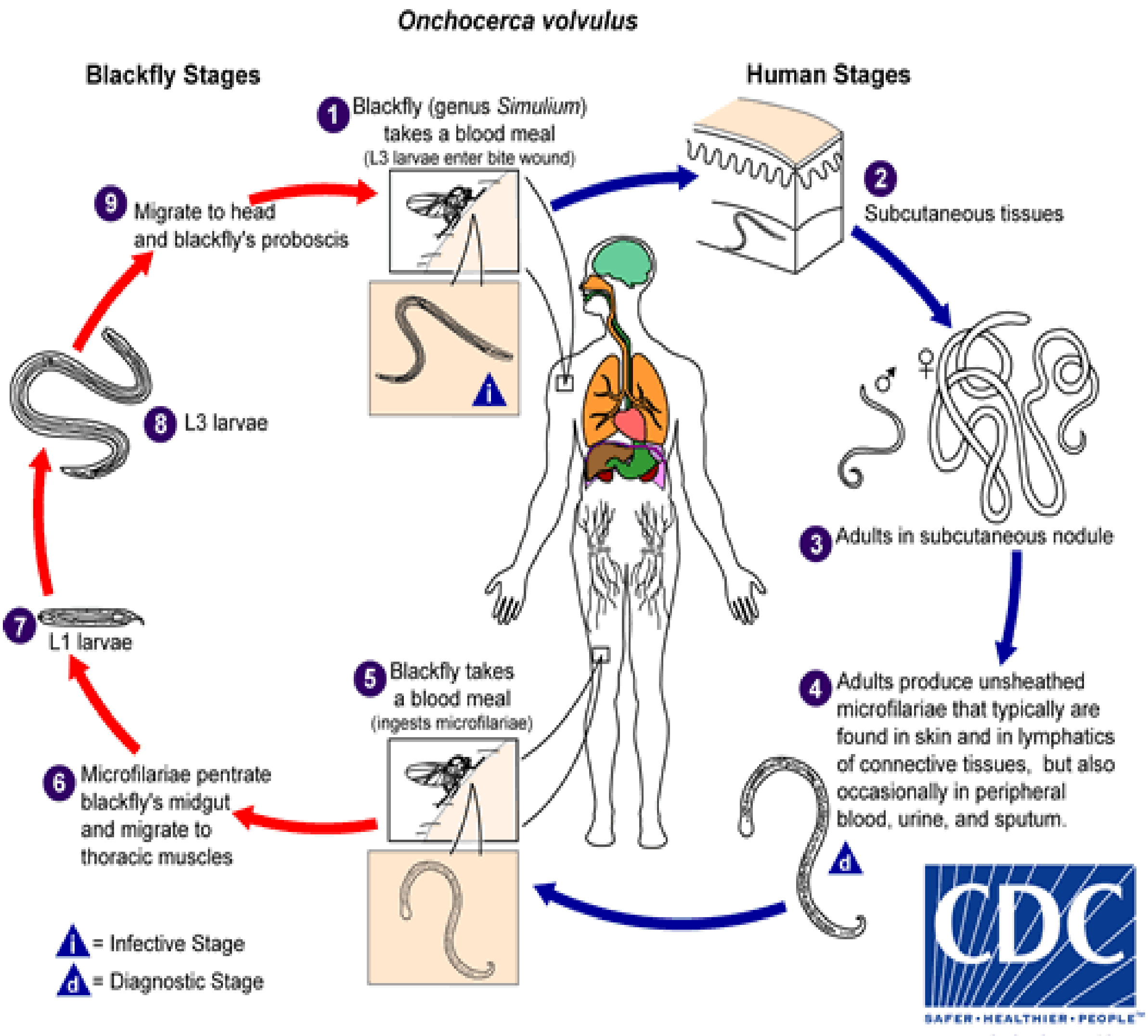

Onchocerca volvulus
Blackfly Stages
Human Stages
1 Blackfly (genus Simulium) takes a blood meal (L3 larvae enter bite wound)
2 Subcutaneous tissues
3 Adults in subcutaneous nodule
♂ ♀
4 Adults produce unsheathed microfilariae that typically are found in skin and in lymphatics of connective tissues, but also occasionally in peripheral blood, urine, and sputum.
5 Blackfly takes a blood meal (ingests microfilariae)
6 Microfilariae pentrate blackfly's midgut and migrate to thoracic muscles
7 L1 larvae
8 L3 larvae
9 Migrate to head and blackfly's proboscis
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER · HEALTHIER · PEOPLE™

Figure 1.1. The life cycle of *O. volvulus* (Centers for Disease Control, 2016).

The microfilariae are unsheathed, can live up to two years and migrate to the skin, eyes, and the lymphatics of connective tissues in the human host (World Health Organization, 2019) (Figure 1.1). In the intermediate host (blackfly), the microfilariae are ingested by a blackfly during a blood meal migrates from the blackfly's midgut through the hemocoel to the thoracic muscles where they develop into first-stage larvae and subsequently into third-stage infective larvae. The third-stage infective larvae migrate to the blackfly's proboscis and can infect another human host when the fly takes a blood meal (Figure 1.1) (Centers for Disease Control, 2019).

1.1.1. Symptoms

The symptoms of onchocerciasis are caused by the microfilariae (detectable in the skin 12–18 months after the initial infection). However, some people do not experience symptoms while infected with *O. volvulus*, as the microfilariae can migrate through the human body without provoking a response from the immune system (Centers for Disease Control, 2019). The adult female worms, which produce thousands of new microfilariae daily, live relatively sheltered from the human immune response in fibrous nodules under the skin and sometimes near muscles and joints. These nodules are formed around the adult worms as part of the interaction between the parasite and its human host. (Centers for Disease Control, 2019). The produced microfilariae move around the human host in the subcutaneous tissue and induce intense inflammatory responses when they die (World Health Organization, 2019). In addition to visual impairment and blindness, the symptoms include debilitating and disfiguring skin disease with depigmentation, severe unrelenting itching with inflammation which results into long-term damage to the skin, and nodules under the skin. The inflammation caused by microfilariae that die in the eye results initially in reversible lesions on the cornea that without treatment progress to permanent clouding of the cornea, and eventual blindness. There can also be inflammation of the optic nerve resulting in vision loss, particularly peripheral vision, and eventually blindness (Centers for Disease Control, 2019). Increasing evidence have also associated high *O. volvulus* infection with different forms of epilepsy and nodding syndrome (a condition in the epilepsy spectrum) in onchocerciasis foci (Chesnais et al., 2018).

1.1.2. *O. volvulus* genomes

The genomes of the parasitic nematode causing onchocerciasis, *O. volvulus,* include the nuclear genome, the mitochondrial genome, and the genome of an intracellular bacterial endosymbiont of the genus *Wolbachia*. The nuclear genome size is approximately 97 Mb and is divided into three autosomes (OM1, OM3, and OM4) and the X-Y sex chromosomes (OM2 and OM5; OM5 is part of the X chromosome and represents the homologous sequence that matches to the Y chromosome and allows pairing during meiosis). OM1 comprises two contigs – OM1a and OM1b – with a sequence gap of at least 50 kb (Cotton et al., 2016). The nuclear genome contains at least one repeat sequence family of a 150 bp repeat known as O-150, which is arranged in tandem arrays and appears subject to concerted evolution (Arnheim, 1983). The mitochondrial genome is compact and 13,747 bp in length (Crainey et al., 2016, Keddie et al., 1998). The estimated *Wolbachia g*enome size is 1.1 Mb (Choi et al., 2016, Klasson et al., 2008) and may play a role in the parasite's fecundity (Hoerauf et al., 1999). Cotton et al. described a total of 12,143 predicted protein-coding genes in the *O. volvulus* genome, the majority (~91%) of which had orthologues in other nematodes and ~9% (1,173) being *O. volvulus*-specific, with little or no homology to genes annotated in other helminths (Cotton et al., 2016).

I will give a brief historical account of the earlier work that described the evolutionary history andgenetic variation in *O. volvulus* genome that was not based on whole genome sequencing, and which suffered from small sample size and strong ascertainment bias. Older work described genetic variation in *O. volvulus* based on markers such as O-150 and other repeat families as well as mitochondrial markers. The nuclear genome of *O. volvulus* is very compact with relatively little non-coding DNA compared to most other eukaryotic organisms. A proportion of non-coding sequences contains two families of non-coding repeated sequences that are usually used as DNA probes to identify *O. volvulus* in genomic screens (Meredith et al., 1991, Zimmerman et al., 1993). One of which is the O-150 repeat family, which is specific to parasites in the genus *Onchocerca,* is the best characterised of these tandemly repeated sequences, and represent roughly 1% of the total *O. volvulus* nuclear genome (Erttmann et al., 1990, Meredith et al., 1989). The O-150 repeat family was a very useful tool for the classification of the two strains of *O. volvulus* in West Africa (rainforest and savanna bioclimes) as the location of the parasites correlated with their pattern of O-150 sequence clustering. Similarly, the O-150 repeat family was used to examine the relationship among different *O. volvulus* populations (Zimmerman et al., 1994). However, the early attempts to use O-150 to correlate genetic variation with ecological parameters (forest vs savannah) and disease outcome (blinding vs non-blinding) proved unreliable (Kron and Ali, 1993). An attempt to examine the level of genetic diversity in the mitochondrial genome of *O. volvulus* using a combination of PCR-RFLP and direct sequencing of the hypervariable AT domain of the mitochondrial genome revealed a very limited genetic variation in the mitochondrial genome based on 11 individual parasites examined from East and West Africa (Zeng and Donelson, 1992, Zimmerman et al., 1994). Unnasch and Williams (2000) reviewed that the *O. volvulus* genome shows large-scale variation in gene density, GC content, and repeat density, but argued based on limited data that intraspecific variation in both the nuclear and mitochondrial genomes are limited.

One possible explanation for that low level of genetic diversity was that *O. volvulus* suffered a huge genetic bottleneck about 5000 - 8000 years ago: since the members of the genus Onchocerca were generally parasites of ruminant animals, while *O. volvulus* was an obligate parasite of humans, *O. volvulus* likely developed because of a recent host switch from one of the endemic ruminants in Africa to humans (Unnasch and Williams, 2000), which would cause a genetic bottleneck. Also, because the *O. volvulus* lifecycle was characterised by significant population (and mostly density-dependent) bottlenecks at each transmission event (BasaÂ ñez et al. 2009). In contrast and based on whole genome sequencing, Choi et al. show clearly that there is a lot of variation in *O. volvulus* genome studied across a large geographic range including samples from Ecuador and several countries in Africa but was also based on very few individuals. The samples were geographically diverse but there was generally only a single individual per sample site, so there were no conclusions concerning how the variation was distributed within and between populations (Choi et al., 2016). The study of Cotton et al. (2016) additionally presents evidence for extensive genetic variation. However, there was a relative lack of information on how that variation was distributed at the population level but also, and more importantly, how that variation was distributed within the genome (Cotton et al., 2016). These more recent, genome-level studies, point to little genetic heterogeneity in the parasite and highlight the need to expand sample size (both in terms of numbers of individuals and in terms of the geographic range of sampling) to understand genetic variation in the genome of *O. volvulus*.

1.1.3. Treatment

The widely used treatment for onchocerciasis is an oral drug called ivermectin (Mectizan^{TM}), from the class macrocyclic lactones (ML), which is given to all members of a community irrespective of their infection status for the life span of the adult worms (approximately, 15 years). Ivermectin was discovered in 1979 and has been shown to have antiparasitic activity against a broad range of nematodes and arthropods (Geary, 2005). Ivermectin has been in use for onchocerciasis treatment since 1988 and has two effects on two discrete stages of the life cycle of *O. volvulus*: the microfilariae and the adult worms: (i) an acute microfilaricidal effect that results in the rapid and almost complete removal of microfilariae from the skin within days to weeks after treatment (Azis, 1982), and (ii) a sustained anti-fecundity effect that results in prolonged but temporary inhibition of the release of new microfilariae from adult female worms into the skin for approximately three to six months(Awadzi et al., 2004c, Duke et al., 1991, Gardon et al., 2002, Grant, 2000, Kläger et al., 1996, Plaisier et al., 1995). Because of this suppression of fertility, ivermectin reduced transmission and the overall incidence of onchocerciasis by reducing the number of microfilariae available for uptake by feeding blackflies (Figure 1.1 stage 5). It also prevents or possibly revert pathology by removing microfilariae from the skin and eyes and delaying repopulation of these tissues with new microfilariae. The recommendation for mass drug administration (MDA) is 15 – 17 years of single annual (or biannual) treatment to interrupt transmission (Turner et al., 2013), past the reproductive lifespan of the adult worms, because the adults survive many years despite treatment and will resume microfilariae production if treatment is stopped.

Clinical trials and subsequent field experiences have shown that ivermectin is a rapidly effective, well-tolerated, single dose microfilaricide, which causes little or no Mazzotti reaction (that is, the severe inflammatory response from the immunological reaction of the body to the death of the microfilariae) (Geary, 2005). Apart from its effectiveness in disrupting transmission of onchocerciasis, it has the secondary effect of reducing intestinal helminthiases in humans (Akogun et al., 2000). This excellent profile makes it suitable for MDA.

1.1.4. Elimination goals

The worldwide burden and transmission of onchocerciasis has been reduced considerably because of successful disease control programs led by the World Health Organisation

(WHO), which were based on vector control and MDA of ivermectin to the affected communities in the Americas and Africa (Centers for Disease Control, 2019, World Health Organization, 2019). For example, in the Americas, four countries have been verified by the WHO as free from onchocerciasis: Colombia, Ecuador, Mexico, and Guatemala (Centers for Disease Control, 2019) and the burden of visual impairment and blindness has been reduced in most of the West African regions affected by onchocerciasis (Murdoch et al., 2002). Since its introduction, ivermectin has been the principal component of control and elimination programs. These programs include the Onchocerciasis Control Programme (OCP) (from 1989-2002; prior 15 years of vector control measures only), the Onchocerciasis Elimination Program for the Americas (OEPA) (1992-2007-2015), with biannual ivermectin treatment (and quarterly in some places), the African Programme for Onchocerciasis Control (APOC) (1995-2009-2015), primarily with single ivermectin treatment annually, and the Expanded Special Project for the Elimination of Neglected Tropical Diseases in Africa (ESPEN) (2017-2025), set up to cover the five preventive chemotherapy NTDs (World Health Organization, 2019).

As noted earlier, MDA of ivermectin and vector control have been successful for controlling onchocerciasis as a public health problem in many foci; however, elimination goals are more stringent and more difficult to achieve (Winnen et al., 2002). There are questions as to whether elimination goals are achievable based on ivermectin MDA alone (Winnen et al., 2002). In a conference about onchocerciasis held at The Carter Center, in Atlanta GA, USA in 2002, it was concluded that eradication of onchocerciasis was not feasible in the hyper- and mesoendemic foci (that is, where 20–40% and >40% of adults respectively have subcutaneous nodules) in Africa with MDA of ivermectin only (Dadzie et al., 2003) and that there was an urgent need to identify the challenges facing elimination objectives. One of the challenges identified was the potential emergence of resistance or sub-optimal response to ivermectin because of its repeated use in preventive chemotherapy (Lazdins-Helds et al., 2003).

Although, resistance to ivermectin in human population infected with *O. volvulus* have not been unequivocally shown, several reports have indicated persistent microfilaridermia, with some patients having high microfilarial counts in the skin, in Africa (Ali, et al. 2002; Awadzi, et al. 2004; LeAnne, 2006 and Churcher, 2009). In 2004, *O. volvulus* were identified in Ghana that exhibited an ivermectin response phenotype termed sub-optimal response (SOR) (Awadzi, et al. 2004c). These SOR parasites were characterised by the presence of live stretched microfilariae in the uteri of the adult worms 90 days after treatment and were associated with repopulation of the skin with microfilariae earlier/more extensively than expected based on prior data. Awadzi et al. (2004c) observed reduction in the post-treatment embryostatic effect of the drug and reappearance of microfilariae in the skin within 2-3 months after multiple doses of ivermectin, even though they observed that the short-term microfilaricidal effect was not affected. From their study, Awadzi et al. confirmed SOR in parasites from people in West Africa collected at different days of ivermectin treatment based on skin microfilarial counts, embryogram data taken at 90 days after treatment and the results of fly-feeding experiments (Awadzi et al., 2004c). In another 30-month follow up study, Awadzi and his collaborators reported persistent significant microfilaremia after several rounds of ivermectin treatment in the worms. Based on embryograms of adult female *O. volvulus,* there were adult worms that responded poorly to repeated doses of ivermectin, that is, they resumed production of skin microfilariae 2-3 months after ivermectin treatment (Awadzi et al., 2004b).

This posed a question on the effectiveness of the embryostatic properties of ivermectin on the reproduction of adult female *O. volvulus.* One important limitation of work on *O. volvulus* is that fertility of a single worm cannot be tracked. To quantify a particular worm’s fertility/fecundity, the worm must be surgically removed and taken apart. In some other nematode parasites (for example, *Schisostoma mansoni),* the complete life cycle can be maintained in the laboratory (Valentim et al., 2013)). However, there are no analogous methods for in vitro maintenance, or for doing controlled crosses or tracking the response of an individual worm to a drug in *O. volvulus*. Thus, validating a hypothesis for mechanism of SOR using the embryogram data is extremely challenging.

The work of Awadzi et al. was controversial and alternative explanations for the variation in the timing and scale of microfilarial reappearance in the skin were proposed by several authors. In response, Churcher et al. (2009) modelled the rate at which *O. volvulus* microfilariae repopulate the skin, starting with the observation that the timing of microfilarial reappearance in the skin following ivermectin treatment does indeed vary considerably between hosts. Churcher et al. suggested that some of the variability may be due to the limitations of the skin snip technique in estimating the true host's microfilarial load (for example, small area of skin sampled and low number of samples per individual). Using an individual-based onchocerciasis mathematical model, they quantified the variability in the post-treatment skin microfilarial repopulation rates among hosts (Figure 1.2) and used the model to generate testable hypotheses to identify whether the unusual rates of skin repopulation by microfilariae was a result of low treatment coverage or decreased ivermectin efficacy. The modelling established that distribution of microfilarial repopulation rates (that is, variation in microfilariae repopulation rates) was primarily affected by the limitations of the skin-snipping method for estimating parasite load, but that time of reappearance was also a sensitive indicator of emerging SOR to ivermectin in *O. volvulus* (Churcher et al., 2009) rather than treatment coverage, that is, modelling support SOR as a likely contributor to variation in the timing of skin repopulation. It is also worth noting that the best predictor of post-treatment microfilariae counts is the count at the time of treatment, that is, a person with a high pre-treatment count will have faster recovery of microfilariae in the skin to a higher level than a person with a low pre-treatment count, and that the skin microfilarial count is the product of a population of adult worms in an individual, that is, it is not a direct measure of individual worm fertility.

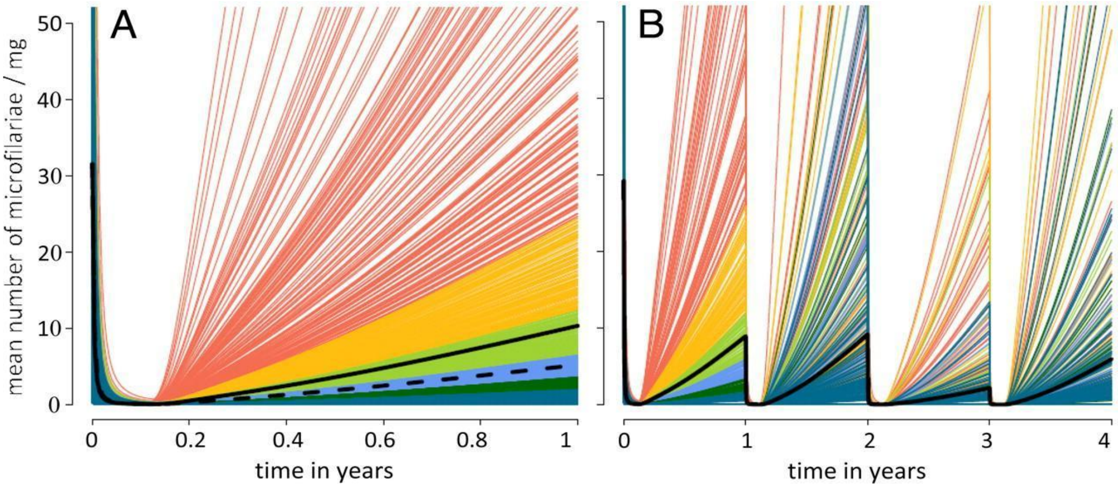


Figure 1.2. The variability in post treatment microfilarial repopulation rates among hosts after (A) their first ivermectin treatment (sample size = 1369) and (B) their first four rounds of ivermectin treatment (sample size = 534). Figure adapted from Churcher et al. (2009).

Individual lines represented the modelled dynamics for each host in the dataset. Lines in both panels are coloured according to the cumulative distribution of responses seen at year 1: lowest 50% (dark blue), 50–60% (dark green), 60–70% (light blue), 70–80% (light green), 80–90% (yellow), and 90–100% (red). The solid thick black line shows the arithmetic mean of all the individual responses (Churcher et al., 2009). The figure indicates that the microfilariae load in the skin decreased at a high rate but suddenly increased and repopulated immediately.

A follow-up study by Osei-Atweneboana et al. (2011) investigated the reproductive response of female worms to multiple treatments with ivermectin and reported evidence that adult female worms were non-responsive or resistant to the anti-fecundity effects of multiple treatments with ivermectin. They conducted this study in 10 endemic communities in Ghana and classified the adult female worms into three categories based on embryogram phenotype and the rate of skin microfilariae repopulation data. The embryograms were taken at day 90 post-infection. The *first category (good responders)* had the expected adult female response to ivermectin; that is, there was complete halt of embryogenesis with barely any skin microfilariae repopulation. The *second category (moderate responders)* had partial adult female response to ivermectin because there are adult worms with microfilariae (and other stages) present in utero at day 90 when all worms should have no viable embryos, and therefore, there existed low to moderate skin microfilariae repopulation. The *third category (poor responders)* have large numbers of live embryos and microfilariae at day 90 because the adult female worms retained their ability to produce microfilariae and the rate of skin microfilariae repopulation was very high (Osei-Atweneboana et al., 2011). The observation of this range of ivermectin response in *O. volvulus* led to the hypothesis that the response phenotype has a genetic basis, and provoked further research in the genetic profiles of the individual worms obtained from categories of phenotypes (good responders, moderate responders, and poor responders) to link the phenotypic poor responses to ivermectin treatment with parasite molecular genetic markers to confirm drug resistance.

In response to this, further research was requested to understand the impact of emerging resistance on control/elimination objectives, to quantify the probability of resistance emerging and spreading within and across geographical areas, and to develop tools and strategies for detecting resistance, and develop strategies to reduce the probability of emerging resistance and to mitigate its impact (World Health Organization and UNICEF, 2018).

First, there is a need to understand what is informing variation between ivermectin resistant and susceptible populations (that is, how did sub-optimal responder phenotype come about). Maybe it was because of selection due to drug use or of standing genetic variation in the original population (That is, how the population was made up by itself). Although there has been some debate regarding the existence of SOR to ivermectin in *O. volvulus* (Burnham, 2007; Cupp et al. 2007; Mackenzie, 2007; Remme et al. 2007; Hotez, 2007), modelling of SOR, using individual-patient data on the rate of skin repopulation by microfilariae following treatment in communities with different histories of control (Frempong et al. 2016; Churcher et al. 2009) has provided support for the conclusion that the early reappearance of microfilariae in the skin that defines SOR is most likely due to a decreased susceptibility of the parasite to ivermectin's anti-fecundity effect.

However, from the first genome-wide analysis of ivermectin response by *O. volvulus*, Doyle et al. (2017) suggested that genetic drift had created genetic differentiation between different *O. volvulus* populations before initiation of ivermectin treatment. They observed a strong, genome-wide genetic differentiation in the NLT populations (worms exposed to a single experimentally controlled round of ivermectin treatment) between Ghana and Cameroon (two endemic foci in Africa). This suggested that the standing genetic variation from which SOR was selected varied significantly between Ghana and Cameroon and may be the reason why there was no consonance between Ghana SOR and Cameroon SOR populations. They also observed strong population structure between Cameroon and Ghana worms which may be because the *O. volvulus* lifecycle was characterised by significant population (and mostly density-dependent) bottlenecks at each transmission event. At such, only a minute percentage of microfilariae under the skin of human hosts are ingested by blackflies, and very few resulting infective larvae are subsequently transmitted to humans and establish as adult worms. Such repeated bottleneck would increase the severity of genetic drift by strongly enhancing the stochastic processes that generate genetic diversity between *O. volvulus* populations, independent of ivermectin treatment. This further, implies that subsequent soft selection of SOR genotypes observed from these genetically distinct populations, from their study, led to SOR populations that are genetically distinct despite their phenotypic similarity.

For selection to occur, there is a need for sufficient generations to have elapsed for the good responder worms to die and be replaced preferentially by SOR worms. This is very unlikely to have happened in *O. volvulus* since 1988 (when CDTI started), even if there were strong pressure from it. Rather the selection pressure observed were almost certainly weak as identified in the study of ivermectin response by *O. volvulus* by Doyle et al. (2017) and Hedtke et al. (2017). When compared with other gastrointestinal nematodes where ivermectin is routinely used (mostly every 3 months), ivermectin is being used annually in *O. volvulus* and one of its roles is to cause temporary infertility in the adult worm for a period of 3 – 6 months (Awadzi et al., 2004c, Duke et al., 1991) (which will increase generation time in *O. volvulus*). Hence, selection pressure would be more intense in other gastrointestinal nematodes when compared to *O. volvulus*. In other gastrointestinal nematodes, phenotype measure is qualitative (for example, survival vs death), while it is quantitative in *O. volvulus* (that is, resistant and susceptible: about how much microfilariae are being produced by adult over time) (Azis, 1982). Further to this, there was a huge population bottleneck between 5000-8000 years ago, as a result of the host switch of *O. volvulus* into humans, which left a significant linkage disequilibrium in the genome (this is about 2000 or 3000 generations of *O. volvulus*). The blocks of LD observed in chapter three of this study indicate the case of bottlenecks. Selection due to drug would not leave such a strong LD in the genome. This strongly suggests that there have been other forms of selection (that is not due to ivermectin, but simply as part of the worms' evolutionary history) that left signatures of LD.

1.2.**Candidate gene approaches.**

Initial studies took a candidate gene approach to investigate the genetic basis of SOR in *O. volvulus*. Analyses were carried out on genes chosen based on specific hypotheses concerning mechanisms of resistance to anthelmintic compounds (Ardelli et al., 2006, Ardelli et al., 2005, Bourguinat et al., 2008, Nana-Djeunga et al., 2012, Osei-Atweneboana et al., 2012). As reviewed by Doyle and Cotton (2019), a candidate gene approach requires an understanding of the biological effects of the drug in question and of the physiological response of the parasite to it. Genes encoding proteins that are involved in these processes become candidates for analysis. An investigation of genetic and biochemical differences between susceptible and resistant parasites is then undertaken to obtain circumstantial evidence for a role for the pharmacologically relevant proteins in conferring resistance. Further functional studies are carried out to prove a causal relationship between a mutation in the candidate gene and the resistance phenotype (Gilleard and Beech, 2007).

In previous helminth studies, it was hypothesized that with intensive use of ivermectin and drug selection response, mutation can occur in several candidates ivermectin response genes (chosen for analysis based on specific hypotheses concerning mechanisms of resistance to the acute effects of ivermectin in them). For example, the P-glycoprotein protein encoding genes and, in the genes, encoding glutamate-gated or ^-aminobutyric acid (GABA)-gated chloride ions channels, leading to ivermectin resistance in both intestinal helminths and arthropods (Currie, et al. 2004; Griffin, et al. 2005; Prichard, 2005). Xu, et al. (1998) discovered that the restriction patterns of P-glycoprotein homologues from *Haemonchus contortus* differs between ivermectin-sensitive and ivermectin-resistant strains, with its increased expression in the resistant strains. In other parasitic nematodes, such as *Haemonchus contortus* and *Cooperia oncophora,* and in free-living nematodes such as *Caenorhabditis elegans*, ivermectin resistance selected for specific alleles of membrane transport genes such as P-glycoprotein (Blackhall et al., 1998, Le Jambre et al., 1999, Xu et al., 1998), beta-tubulin, and glutamate-gated chloride channel genes (Blackhall et al., 1998, Dent et al., 2000). Allele frequency change in those candidates ivermectin response genes has also been demonstrated in *O. volvulus* populations when sampled before and after several rounds of ivermectin treatment (Ardelli et al., 2006, Ardelli et al., 2005, Bourguinat et al., 2008, Eng and Prichard, 2005, Huang and Prichard, 1999, Nana-Djeunga et al., 2012, Osei-Atweneboana et al., 2012).

For example, a comparison of the genetic polymorphisms in populations of the worms from ivermectin treated and untreated patients gave evidence of genetic selection in them (Eng and Prichard, 2005). This selection was found to occur in ABC transporter gene which functions as an energy-dependent efflux pump, similar to P-glycoprotein (Ardelli, et al. 2006). Huang and Prichard (1999) also discovered that the restriction patterns of P-glycoprotein homologue cloned from *O. volvulus* was expressed differently between the larval and adult stages, and further studies suggested high selection pressure on that site of the parasite and might be consistent with ivermectin resistance at that point (Bermadette, et al., 2005; Eng and Prichard, 2005). Also, another study on tubulin gene gave an evidence of genetic selection on the gene (Eng and Prichard, 2005). However, the relationship between the genetic polymorphisms and suboptimal clinical response to ivermectin was not yet determined and whether those genetic changes are an indicator of developing ivermectin resistance or not (Bernadette, et al. 2005; Gomez-Priego, et al. 2005; LeAnne, 2006).

Eng and Prichard (2005), while comparing the genetic polymorphisms in 16 genes (six genes which were previously suggested in other nematodes as having possible association with ivermectin resistance and ten genes which were included as control genes) in populations of the worms from ivermectin treated and untreated patients, presented evidence of selection in an ABC transporter gene which functions as an energy-dependent efflux pump, similar to P-glycoprotein, and β-tubulin (Eng and Prichard, 2005).

Bourguinat et al. (2007) suggested in their study on the genetic selection of low fertile *O. volvulus* by ivermectin treatment, that ivermectin-resistance selection in the parasite is associated with a lower reproductive rate in the female parasites. In the study, the genetic changes in parasites obtained from the same person prior to and after several levels of ivermectin exposure (that is, after administering variation in the dose of ivermectin treatment) were monitored, and they observed different genotype frequencies in the following genes: β-tubulin, heat shock protein 60, and acidic ribosomal protein (Bourguinat et al., 2007). Later, in 2008, they again argued that P-glycoprotein-like protein is a possible genetic marker for SOR selection in the parasite because three of the six polymorphic positions found in the P-glycoprotein-like protein amplicon showed significant selection after 4 times per annum treatment with ivermectin (in a total of 13 ivermectin treatments) in female worms (Bourguinat et al., 2008). Similarly, Nana-Djeunga et al. investigated the four single nucleotide polymorphisms (SNPs) occurring in the β-tubulin gene of *O. volvulus* adult worms collected from the same individuals before and after three years of ivermectin exposure and observed changes in genotype frequencies in the *O. volvulus* β-tubulin gene associated with ivermectin treatments (Nana-Djeunga et al., 2012). These studies have focused on the role of candidate genes using simplistic models of single gene selection. However, evidence has suggested that the evolution of macrocyclic lactone (ML) resistance in nematodes is more complex and that there are several, perhaps many, different genetic mechanisms for selection involved (that is, soft selection on a polygenic trait is acting on ML resistance in nematodes, including *O. volvulus*) (Bourguinat et al., 2015, Choi et al., 2017, Doyle et al., 2017, Hedtke et al., 2017). As a result, genome-wide association studies (GWAS) could replace candidate gene approaches in studies regarding genetics of drug resistance in helminth parasites (Doyle et al., 2017, Doyle and Cotton, 2019).

1.3. Soft selection and difficulties associated with candidate gene approaches and the advantage of GWAS.

Candidate gene approaches face a number of challenges, particularly in parasitic nematodes. The challenges of using candidate gene approach in *O. volvulus* are that (1) functional validation steps cannot be done as detailed in Geary (2005). Functional validation is difficult in parasitic nematodes regardless of how the candidate gene came to be proposed. The strength of GWAS is that it is unbiased, and that, provided one takes appropriate steps to correct for population structure and multiple testing, it is less likely to result in spurious associations that are frequently observed in single gene candidate associations (which rarely take population structure or multiple testing into account), and (2) because the evidence suggests soft selection on a polygenic trait is acting on ML resistance in nematodes, including *O. volvulus* (Doyle et al., 2017, Doyle and Cotton, 2019) and GWAS gets around the problem of the candidate gene approach in terms of being able to detect and describe multiple genes involved in soft selection and QTLs (Gilleard, 2006). Single candidate gene approaches (or even those that may allow for two or three genes) assume hard selection whereas GWAS makes no assumption as to the mode of selection, although selection mode can impact GWAS results

GWAS has been used extensively to discover potential mechanisms and pathways that underlie diseases and drug responses (Bourguinat et al., 2015, Choi et al., 2017, Doyle et al., 2017, Gudbjartsson et al., 2015, Manolio, 2010, Manolio et al., 2008). If the trait of interest is a drug resistance, then the locus/loci involved should be under selection in populations exposed to drug. In principle, there are two ways in which selection could bring about changes in allele frequencies at those loci. GWAS offers a method by which those loci can be discovered without any information or prior assumptions about mechanisms. Although, caution is needed. For example, pleiotropic interactions can likely induce false positive signals in GWAS. In the wild, this is certainly the case if an individual fitness is affected.

First, in a "hard" selective sweep, a rare beneficial mutation or resistance-conferring allele arises (usually, during the period of initial application of the selective pressure, for example, a drug) and increases in frequency rapidly, thereby drastically reducing genetic variation in the population (Smith and Haigh, 1974). In a hard sweep, the lineages in the sample that carry the resistance-conferring allele (or mutation) coalesce more recently than the onset of positive selection, that is, the point in time when it first became advantageous to carry the allele (or at the time of drug exposure). Figure 1.3A shows an example of the genealogy of resistance-conferring alleles at a selected site after a typical hard sweep. For example, if ivermectin response is a result of hard selection, all resistance-conferring alleles in the sample will arise from a single mutation (depicted by x in the figure) and coalesce after the onset of positive selection (drug exposure) (Figure 1.3A); that is, they will be monophyletic (Messer and Petrov, 2013).

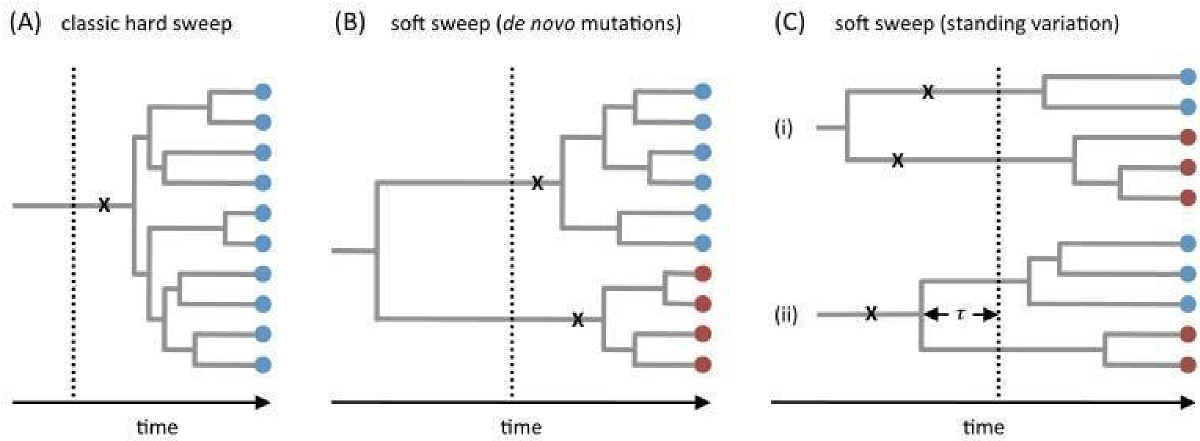

(A) classic hard sweep
x
time
(B) soft sweep (de novo mutations)
x
x
time
(C) soft sweep (standing variation)
(i)
x
x
(ii)
x
τ
time

Figure 1.3. Definition of hard and soft selective sweeps. **Image adapted from Messer and Petrov (2013).**

(A) In a hard sweep, all adaptive alleles in the sample arise from a single mutation (depicted by x) and coalesce after the onset of positive selection (dotted line). Note that even if the mutation had arisen prior to the onset of positive selection and was present as standing genetic variation, this would still be considered a hard sweep if only a single lineage is ultimately present in the sample. (B) In a soft sweep from recurrent *de novo* mutations, the adaptive alleles in the sample arose from at least two independent mutation events after the onset of positive selection and the lineages coalescence prior to the onset of positive selection. (C) In a soft sweep from the standing genetic variation, adaptive alleles were already present at the onset of positive selection. The different lineages in a population sample can originate from independent mutation events (i) or from a single mutation that reached some frequency prior to the onset of positive selection, such that several copies present at that time then swept through the population (ii). In this latter case, the population genetic signatures of the sweep will depend on the time τ between coalescence and onset of positive selection. If τ is short, the sweep will appear like a hard sweep, whereas when τ is large, it will be like a soft sweep from several *de novo* mutations (Messer and Petrov, 2013).

Moreover, since it takes several generations to emerge, recombination is rare around the allele under selection. The surrounding region in the genome will be in strong linkage disequilibrium (LD) with the resistance-conferring allele. Linkage disequilibrium refers to the non-random association of alleles at two or more loci in a general population (Bates, 2005, Gusella et al., 1983, Qanbari, 2020). When loci are in LD, the frequency of association of alleles is higher or lower than what would be expected if the loci were independent and associated randomly (Slatkin, 2008). Aside selection or drift, LD between two alleles is determined by two major factors: the physical distance between the two loci on a chromosome (linkage) and the recombination rate in the region included between the two loci (Goode, 2011, Gusella et al., 1983, Qanbari et al., 2010). Recombination erodes LD over generations, but the rate at which that occurs depends on three parameters: how many generations since the mutation arose or the hard selection was applied, how far away a site is from the locus under selection, and how large the population is. Linkage disequilibrium is maintained for many generations between alleles close to the locus under selection, but for fewer generations between alleles much further away from the locus under selection. The number of generations is the product of time and population size (Messer and Petrov, 2013). Therefore, in the case of *O. volvulus,* selection of resistance-conferring alleles at the resistance locus would cause regions of strong LD around resistance-conferring alleles or increased genetic differentiation (measured, for example, by Wright's FST statistic) between GR and SOR worms. That strong signal of genetic differentiation, with the same allele in the same "hitchhiking" surrounding genomic environment, will be present in all survivors of treatment in that population (SOR in this case) and their progeny provided they contain the resistance conferring allele (Gilleard and Beech, 2007, Stephan et al., 2006).

A typical example of a hard selective sweep was observed in the human population in the independent selection in Europe and Asia for persistence of lactase expression in adults (Tishkoff et al., 2007). Lactase is the enzyme that catalyses the first step in lactose metabolism, and, in most mammals, expression of lactase is much reduced or absent in adults, resulting in lactose intolerance. The evolution of lactase expression in adults followed the domestication of dairy animals and inclusion of dairy products in the human diet. The same changes in lactase expression evolved independently in all human populations that utilise dairy products, leaving strong signatures of hard selection in the genomic region surrounding the lactase gene in the genomes of those populations

(Tishkoff et al., 2007). Other well-known examples include the evolution of pesticide resistance in insects (Daborn et al., 2001), colour patterns in beach mice (Hoekstra et al., 2006), and freshwater adaptation in sticklebacks (Colosimo et al., 2005).

Second, a "soft" selection sweep: this describes the situation where multiple resistance-conferring alleles at the same locus sweep through the population at the same time (Hermisson and Pennings, 2005). Soft sweeps usually arise in two ways as demonstrated in Figure 1.3 B&C: (a) the mutations can arise *de novo* after the onset of positive selection (that is, the point in time when it first became advantageous to carry the allele or at the time of drug exposure) (Figure 1.3B) or (b) the mutations were already present previously as standing genetic variation (Figure 1.3C, top row). Finally, a situation where the resistance-conferring allele arose only once but reached some frequency prior to the onset of positive selection (via genetic drift, for example) and several copies then swept through the population, is still considered a soft sweep if the lineages coalesce prior to the onset of positive selection (Figure 1.3C, bottom row) (Messer and Petrov, 2013). In a soft sweep, lineages collapse into more than one cluster and several haplotypes can be frequent in the population at the adaptive locus. Such a mutation may be present on several genomic backgrounds so that when it rapidly increases in frequency, it does not erase all genetic variation in the population (Hermisson and Pennings, 2005). Diversity is thus not necessarily reduced and deviations in the frequency distributions of neighbouring neutral polymorphisms are typically very weak compared to hard sweeps.

In the case of *O. volvulus* and ivermectin response, pre-existing alleles (as a result of mutations) that have been present for a significant period prior to ivermectin exposure are proposed to have accumulated and undergone recombination to generate a pool of haplotypes (or QTLs) that all contain the resistance-conferring mutations (Gilleard, 2006). This feature of a QTL that is produced by soft selection means that the alleles under selection generally occur on many different haplotypes, or genetic backgrounds, so there is usually only weak, and variable, LD between the allele under selection and its immediate genomic neighbourhood.

Soft selective sweeps may be more common in eukaryotes than previously recognized, particularly for organisms with large census population sizes (Messer and Petrov, 2013). For example, resistance to organophosphate insecticides in *Drosophila melanogaster* and

Lucilia cuprina involve multiple independent resistance alleles at the acetylcholinesterase and esterase loci, respectively (Claudianos et al., 1999, Daborn and Le Goff, 2004). Soft sweeps have been documented in sticklebacks (Feulner et al., 2013) and beach mice (Domingues et al., 2012). In humans, several cases of selection from standing genomic variation have been reported (Bhatia et al., 2011, Peter et al., 2012, Seixas et al., 2012).

This pattern of selection is also evident in the GWAS of anthelmintic resistance in helminths, including *O. volvulus*, *Dirofilaria immitis* and *Teledorsagia circumcincta* (Bourguinat et al., 2015, Choi et al., 2017, Doyle et al., 2017, Doyle and Cotton, 2019, Gilleard and Beech, 2007, Hedtke et al., 2017). For example, Doyle et al. (2017) reported that soft selective sweeps contribute to loss of drug sensitivity in *O. volvulus* and that ivermectin response does not involve a single mutational event that subsequently rapidly sweeps through all parasite populations. Instead, it probably involves a mixture of pre-existing mutations, recurrent recent mutations, and migration of resistance alleles between populations (Doyle and Cotton, 2019, Gilleard, 2006). Ivermectin resistance is a polygenic, quantitative trait. A good feature of a quantitative trait is that the same phenotype (or trait value) can be achieved by the additive contributions of different alleles at different loci (QTL), such that two individuals within a population may show the same trait value (degree of anthelmintic resistance, phenotype) but have different genotypes (Doyle et al., 2017, Doyle and Cotton, 2019). Anthelmintic resistance can arise because of both modes of selection (hard and soft), and with appropriate analysis, the genetic characteristics of both selections could be differentiated from each other (Redman et al., 2015).

In the context of onchocerciasis elimination, Grant (2000) proposed that selection for SOR of adult worms to ivermectin (that is, an early return to fertility by adult female worms following a single treatment) might be more serious for *O. volvulus* control than selection for microfilariae resistance (that is, a failure to clear microfilariae from the skin at the time of treatment), and thus adult female worms can drive the potential for SOR and should be the focus of study when considering the potential for SOR to ivermectin (Grant, 2000).

1.4. Genome-wide Association Study (GWAS).

A Genome Wide Association Study (GWAS) is the examination of many genetic variants distributed over the entire genome, using high-throughput genotyping technologies of hundreds of thousands of SNPs, across different individuals to determine if any variant is statistically associated with a trait of interest such as drug response (Pearson and Manolio, 2008, Spencer et al., 2009). There are several alternative approaches to GWAS. *First and most common is a case-control study design*: loci in the genome are interrogated for association with a trait using SNPs by comparing allele frequencies in case (that is, individuals displaying the phenotype of interest) and controls (individuals drawn from the same population that do not display the phenotype (Manolio, 2010)). *A second approach is a cohort design,* in which extensive baseline information for many individuals is observed to assess the incidence of disease subgroups defined by genetic variants (Weedon et al., 2007). *A third option is the trio design* in which affected case participants and their parents are genotyped, and the frequency with which an allele is transmitted to the affected offspring from heterozygous parents is then estimated (Connolly and Heron, 2015). *Last is multistage design*, in which genome-wide scans are performed on an initial group of case and control participants and then a smaller number of associated SNPs is replicated in a second or third group of case and control participants (Hirschhorn and Daly, 2005). The non-hypothesis-propelled characteristics of GWAS makes it a step beyond candidate gene studies (Pearson and Manolio, 2008). Due to the vast number of tests of association required (at least one per SNP) in GWAS and stringent statistically significant thresholds, there is a need to work with a very large number of samples (Hunter and Kraft, 2007). Unfortunately, the issue of generating large sample size is a typical problem in *O. volvulus.*

GWAS have proven successful in identifying variety of variants associated with common and complex diseases (Hindorff et al., 2009). For example, the findings from large-scale sequencing projects like the 1000 Genomes Project and the Exome Sequencing Project (ESP) have given novel insight into deeper understanding of human genome diversity, the structure and history of the human population and the tools to use in genetic discovery (1000 Genomes Project Consortium, 2012, Fu et al., 2013, Tennessen et al., 2012). For example, Gudbjartsson *et al.* shared the great insights gained from whole genome sequencing of the Icelandic population. A “trio design” which used pedigrees of families that were first identified in a cohort study of the whole population was implemented. Twenty million SNPs and 1.5 million indels were observed. The density and frequency spectra of sequence variants in relation to their functional annotation, gene position, pathway and conservation score were described, and an excess of homozygosity and rare protein-coding variants were also revealed (Gudbjartsson et al., 2015). In addition, Gudbjartsson et al. gave a comprehensive understanding of the basis of using imputation to discover associations between variants in sequence and phenotypes because they were able to impute variants in 104,220 individuals down to a minor allele frequency of 0.1%. This study serves as an example for GWAS and because it demonstrates how GWAS can discover causal relationships that would not be predicted by a candidate gene approach; that is, GWAS is unbiased and facilitates discovery of novel mechanisms (Manolio, 2010, Manolio et al., 2008).

GWAS, that is association study, and genome-wide scan of variation (that is F_{ST} analysis) are both performed with genome-wide resolution but under different frameworks and assumptions. GWAS is statistical test for association between a phenotype and SNP one-at-a-time across the genome. The phenotype could be a binary or continuous trait. F_{ST} is a measure of allele frequency differences, it could be one or a group of markers and two or several populations, either across the genome or at some locations of the genome (Holsinger and Weir, 2009). It is important to note that genome-wide scan is far from perfect, especially when using limited population set, poorly defined phenotypes and improving genomic resources. Both approaches are essential and needed to understand how a drug works (candidate gene approach) and how selection for resistance emerges under natural selection (QTL study) (Doyle and Cotton (2019).

1.5.**Success of GWAS in helminths.**

GWAS has helped in discovering genetic markers predictive for drug response phenotypes, and in discovering plausible mechanisms that underpin variation in drug response in helminths. For example, multiple loci have been identified as correlated with ML resistance in multiple parasitic nematodes, including *D. immitis* (Bourguinat et al., 2015) and *T. circumcincta* (Choi et al., 2017). Studies have indicated that loss of efficacy (LOE) of MLs used as chemoprophylaxis for *D. immitis* infection in dogs has become common in some locations in the USA (Blagburn et al., 2011, Bowman, 2012, Hampshire, 2005, Pulaski et al., 2014). Bourginat et al. (2015) used GWAS to investigate the loci that could be associated with the ML resistance phenotype in this parasite. They compared the whole *D. immitis* genomes of the suspected (Loss Of Efficacy) and the confirmed ML resistant isolates from a controlled efficacy study to the genomes of ML susceptible heartworms. They concluded that ML resistance in *D. immitis* is a polygenic trait (that is, with many genes involved) as many loci showed highly significant differences between pools of susceptible and LOE isolates of *D. immitis* (Bourguinat et al., 2015). An interesting feature of this (and other) GWAS of ML LOE/SOR is that the GWAS generally fail to detect candidate genes proposed in candidate gene association studies. The discrepancy between the large number of candidate genes proposed and QTL defined by GWAS has yet to be resolved. This simply points at the drawbacks of candidate gene approach: Candidate genes usually come from screening assays that identify the very receptor targeted by a molecule. It is not known how this mutation provides additional fitness under the field conditions.

Similarly, to shed more light on the genetic architecture of multiple anthelmintic resistance in parasitic nematodes, Choi et al. (2017) examined multidrug resistant *T. circumcincta*, a major parasite of sheep, using a comprehensive whole-genome analysis. A field-derived, multiresistant genotype of *T. circumcincta* was backcrossed into a partially inbred susceptible genetic background (through repeated backcrossing and drug selection), and genome-wide scans were conducted in the backcrossed progeny and drug-selected second filial generation (F2) populations to identify the major genes responsible for the multidrug resistance. Their findings identified several QTLs that differentiate between the resistant and susceptible populations and may contribute to variation in target site sensitivity, reduced target site expression, and increased drug efflux (Choi et al., 2017). This reinforced the hypothesis that drug resistance in these parasites is a multifactorial quantitative trait rather than a simple discrete Mendelian character. It is interesting to note that the Choi et al. analysis of ML-resistance in *T. circumcincta* is the only example to date of a candidate gene being validated independently by GWAS (the *Tcirc-pgp-9* gene). Possible causes for this discrepancy are discussed at length in Doyle and Cotton (2019).

Another use of GWAS to discover genetic markers predictive for drug response phenotypes in helminths is seen in the work from Tim Anderson's lab on praziquantel resistance in *Schistosoma mansoni*. Valentim et al. (2013) demonstrated how genome sequence data can be leveraged for functional genomic analyses of a biomedically important trait (drug resistance) in a neglected human helminth parasite. They combined GWAS with genetic mapping and functional validation of candidate genes within QTL to identify resistance loci and to determine the molecular basis for species-specific drug action. Using crossed parental parasites differing approximately 500-fold in drug response, they determined drug sensitivity and marker segregation in clonally-derived F2s and identified a single QTL (LOD=31) on chromosome 6. Using RNAi knockdown and biochemical complementation assays, a sulfotransferase was identified as the causative gene (but not the causative allele) and subsequently demonstrated independent origins of loss-of-function mutations in field-derived and laboratory-selected resistant parasites (Valentim et al., 2013). Studies on *S. mansoni* have *a*n advantage over *O. volvulus* research in that its complete life cycle can be maintained in the laboratory and clonal expansion of larval parasites within the snail host is possible. This allows production of thousands of genetically identical single sex parasites, making this organism well suited to linkage mapping methods (Valentim et al., 2013).

Another important example in helminths where GWAS was used to identify novel causal variants was the discovery of the mechanisms that gave rise to benzimidazoles (BZ) resistance in the free-living nematode *C. elegans.* Hahnel et al. (2018) took an unbiased genome-wide mapping approach in the free-living nematode species *C. elegans* to identify the genetic underpinnings of natural resistance to albendazole (ABZ) (the commonly used BZ). In concordance with the known mechanisms of BZ resistance in parasites, they argued that most of the variation in ABZ resistance among wild *C. elegans* strains was caused by variation in the β-tubulin gene *ben-1*. They further identified a novel genomic region that is correlated with ABZ resistance in the *C. elegans* population but independent of *ben-1* and the other β-tubulin loci, suggesting that multiple mechanisms underlie BZ resistance in *C. elegans* (Hahnel et al., 2018). Choi et al. in their study on *T. circumcincta*, also included GWAS for BZ resistance, and showed that the GWAS found the correct isotype-1 β-tubulin locus as the major determinant. Thus, the known “candidate” for BZ resistance was confirmed by the GWAS (Choi et al., 2017). Similarly, GWAS exposed different QTL underlying bleomycin (a medication used to treat cancer) response variation in the recombinant strain of *C. elegans* other than the one identified using linkage mapping approach, suggesting genetic complexity underlying the bleomycin response phenotype in the worm (Brady et al., 2019). An advantage with *C. elegans* is it serves as an excellent model for basic cellular and organismal processes (Corsi et al., 2015) because of its well annotated reference genome (C. elegans Sequencing Consortium*, 1998) and its broad genomic diversity across global populations (Cook et al., 2017).

In the first study to use genome-wide scan of variation using F_{ST} in *O. volvulus*, Doyle et al. (2017) gave insights into the genetic variation that significantly differentiated GR and SOR in adult female worms from Ghana and Cameroon using pooled next generation sequence of worms that varied in ivermectin treatment history and response. They proposed that ivermectin response is a polygenic quantitative trait in which similar molecular pathways influence the extent of ivermectin response in the various parasite populations, and not discrete genes. Doyle *et al.* specified that the variants that differentiated GR and SOR parasites are gathered in about 31 QTLs and contain genes in molecular pathways associated with neurotransmission, development, and stress responses, although additional studies are necessary to validate the putative QTLs identified. They also noted that previously proposed candidate ivermectin SOR genes were largely absent in the regions identified as QTLs differentiating GR and SOR worms (Doyle et al., 2017). An important key finding of this study is that the QTLs were different between Ghana and Cameroon. Thus, this suggests that the genetics of SOR are different between different populations (as would be expected for soft selection on a quantitative trait from standing genetic variation), and that if SOR genetics are going to be studied, then association studies will need to be done on each population. Further, the study lends credence to the hypothesis that ivermectin response in *O. volvulus* is a multi-gene trait under soft selection. However, the study of Doyle et al. was based on limited number of low sequence coverage pool-seq worms, which resulted into stochastic variation in allele detection in the worms. Additional GWAS would need to be done on individual worms and multiple populations, raising the challenges of sequencing cost when studying the genetics of SOR in *O. volvulus*.

1.6.Problems with implementing GWAS or genome-wide scans in *O. volvulus.*

The main "problem" with genome-wide scan in *O. volvulus* is sample size, and there are several factors that all contribute to the need for large sample size. One important factor is that fertility of a single worm cannot be tracked. That is, to quantify a worm's fertility/fecundity, the worm must be surgically removed and dissected to examine the uteri. This is a notable caveat because it is connected to the challenge of working with this parasite in vitro compared to some other helminths (for example, the complete life cycle of *S. mansoni* can be maintained in the laboratory and clonal expansion of larval parasites within the snail host is possible (Valentim et al., 2013)). In contrast, in *O. volvulus* there is no mechanism for doing controlled crosses or for tracking the response of an individual to a drug. Thus, making the validation of SOR mechanism in *O. volvulus* very challenging.

A related and equally important caveat is that nodulectomy, the surgical removal of papular nodules surrounding adult worms from the infected person (Richards et al., 2000), is random with respect to the worm(s) that are removed. That is, there is no way to ensure that the worms that are phenotyped are responsible for the microfilariae in the skin. This suggests that SOR worms may not be included in the sample, so there is the need for a larger sample.

Due to population structure, genome-wide scan needs to be performed on individual worms by populations, which is another major caveat. The first published genome-wide scan of ivermectin response in *O. volvulus* using pooled worms noted that Cameroon worms were different from Ghana, and as a result, the genetics of SOR are different between them (Doyle et al., 2017). Thus, to study SOR genetics the genome-wide scan should be done on single populations following first testing for population structure and assigning individual worms to the correct population. Pool-seq is more economical but leads to uncertainty in estimating allele frequency (and loss of haplotype information), hence the need for a bigger sample. In addition, sequencing single worms instead of pooled worms would increase statistical robustness for each population under study but moving from pool sequencing to single worm sequencing is also limited by the cost of whole genome sequencing of individual worms.

Hedtke et al. (2017) attempted to solve this challenge by performing additional genome-wide scan on genome sequences of 48 individual phenotyped *O. volvulus* form Ghana using Wright's F statistics (F_{ST}) to identify regions with strong association with SOR phenotype. The study revealed several QTLs distinguishing GR and SOR worms from Ghana, but there is a need for a follow up study with larger sample size for validation of the observed QTLs. Unfortunately, additional worms phenotyped from the same population did not have sufficient quality from the genomic DNA extractions and thus additional whole genome sequencing of single worms for genome-wide scan was not possible. DNA quality is clearly an issue that limits sample size. An important practical constraint on moving from pool to single worm sequencing is the frequently low yield and/or poor quality of genomic DNA from adult worms. It is not clear which factors contribute to this poor quality, but they probably include the need to first remove nodules from infected people under field conditions, transport and store the nodules while maintaining a cold chain, digest nodules with collagenase to release the adult worms, and only then preparing genomic DNA. The yield of genomic DNA is low and variable, as is the quality. There is also a variable degree of host (human) DNA contamination that compromises sequencing. An alternative approach would be to validate the putative QTLs using PCR amplification of targeted regions, which is less sensitive to input DNA quantity and quality.

Another caveat is that GWAS requires a vast number of association tests (at least one per SNP) and stringent statistical thresholds to confidently match a genotype with phenotype. And GWAS power also depends on LD between markers and the true QTL. This requires a large number of samples (Hunter and Kraft, 2007, Spencer et al., 2009). In *O. volvulus*, increasing the sample size to compensate for weak LD between a marker and a causative QTL is impractical because of the overall lack of good quality genomic DNA, the cost of genomic sequencing and uneven genome coverage of those worms already sequenced, all of which limit the number of samples available for GWAS (Doyle and Cotton, 2019, Hedtke et al., 2019).

Whole genome sequencing and genotyping of all the identified variants in the genome is almost certainly impractical. However, because the genotypes at nearby markers are usually correlated (that is, they are in LD), it may be possible to scan the genome using a much smaller marker set with only a modest loss of power to detect selection while minimising the amount and quality of DNA that is required. More importantly, the power of a relatively small (compared to the whole genome) group of markers to detect selection can be augmented significantly using methods that predict the genotype at un-typed/missing loci from a set of reference genotypes (known as a ‘reference panel’) by imputation (Nothnagel et al., 2009). To design such studies, it is necessary to have a detailed understanding of the structure and extent of LD across the genome, both to choose suitable reference and genotyping marker sets and to design appropriate methods of statistical analysis (Li et al., 2009).

The development of a complete reference genome of *O. volvulus* opened the door for my study which explored various methodologies to improve the problems facing GWAS of ivermectin response. Even though the set of methodologies employed in this thesis did not solve the problem of lack of statistical power, it has helped in improving the *O. volvulus* reference genome by providing information on the distribution of variation across the genome and the LD interval between an observed QTL and the surrounding SNPs as we move along the chromosome. This will aid in pinning down the location of the causative SNPs faster, understand the mechanism of SOR in the parasite and eventually aid in achieving the timely elimination of onchocerciasis.

1.7.**Linkage Disequilibrium (LD).**

As mentioned earlier, QTLs that are associated with drug response have been identified in *O. volvulus* (Doyle et al., 2017, Hedtke et al., 2017) and association mapping is the most common approach to mapping QTLs that takes advantage of the historic LD to connect phenotypes to genotypes. However, the design and the success of this study depends on a detailed description of the LD across the organism's genome (reviewed in Goddard and Hayes, 2009).

All GWAS rely on LD between genetic markers that differ between phenotypic class and the loci that are under selection. Selection is not the only mechanism for LD; for example, admixing genetically distinct populations through migration creates association between two loci with different allele frequencies even if they are unlinked. LD can also arise due to population stratification and cryptic relationships within a population that result in correlated allelic frequencies (reviewed in Hellwege et al., 2017). Factors, such as inversion, that prevent alleles of genes from undergoing recombination through a chromosomal abnormality also causes LD. Organisms that undergo a high level of inbreeding, as through self-fertilization, may also display significant LD, owing to the reduced opportunities for recombination (Hartl and Clark, 2007).

Information about the genome-wide distribution and extent of LD is critical for:

1. Adjusting estimates of statistical power in GWAS, in which an assayed polymorphism may not itself be contributing to the phenotype under study but is in LD with the causal polymorphism (Goode, 2011).
2. Selecting markers to locate QTLs in a GWAS (Carlson et al., 2004).
3. Estimating how many markers will be needed to achieve acceptable power in genome-wide studies (Meadows et al., 2008); for example, if LD were extensive, the number of markers needed for genome-wide test of association will be

reduced, with low probability of missing the association (International HapMap Consortium, 2005).

4. Designing statistical methods of analysis that make optimal use of the data such as genotype imputation (Wall and Pritchard, 2003a).

5. Studying the evolutionary demographics of a population (García-Gámez et al., 2012, Pavlidis and Alachiotis, 2017); for example, LD can be used to infer changes in a population's effective size (N_e) through generations (Hill, 1981, Sved, 1971). N_e is an important population parameter which helps to explain how populations evolved and expanded, and therefore needs to be incorporated into studies about the genetic architecture underlying complex traits (Reich and Lander, 2001). Populations with smaller N_e have experienced more genetic drift than larger ones, and genetic drift causes LD between alleles at independently segregating loci at a rate inversely proportional to N_e (Wang, 2005). In the same vein, LD is a function of N_e in Sved's model given that the inter-loci recombination fractions are available (that is, N_e and the recombination rate are used to predict LD) (Sved, 1971).

6. Tracing selective sweeps (Biswas and Akey, 2006, Stephan et al., 2006). The signatures of genomic regions under positive selection can be identified by studying the pattern of LD that emerges between SNPs in the vicinity of the target site for positive selection (Pavlidis and Alachiotis, 2017) (Figure 1.3). Upon fixation of the beneficial mutation, elevated levels of LD emerge on each side of the selected site. The high LD levels on the different sides of the selected locus are because a single recombination event allows existing polymorphisms on the same side of the sweep to escape the sweep. On the other hand, polymorphisms that reside on different sides of the selected locus need a minimum of two recombination events to escape the sweep (Messer and Petrov, 2013, Pavlidis and Alachiotis, 2017). Given that recombination events are independent, the level of LD between SNPs that are located on different sides of the positively selected mutation decreases (Pavlidis and Alachiotis, 2017, Doyle and Cotton, 2019).

LD is essential in association studies in which variants can be detected through the presence of association at nearby sites (Wall and Pritchard, 2003b). If most affected individuals in a population share the same mutant allele at a causative locus, it is possible to narrow the genetic interval around the locus by detecting disequilibrium between nearby markers and the disease locus (Altshuler et al. 2008). This approach makes use of the many opportunities for crossovers between markers and the locus during the many generations since the first appearance of the mutation. As a result, there has been a resurgence of interest in LD, owing largely to the belief that association studies offer substantially greater power for mapping common disease genes than do traditional linkage studies, and that LD can offer a shortcut to GWAS as it allows identifying genetic markers that tag the actual causal variants to complex human diseases (Monteiro et al., 2016).

The two most used measures to evaluate LD, for bi-allelic markers, are r^2 and *D'* (Hill and Robertson, 1968). *D'*, as suggested by Lewontin and Kojima, is a normalized *D* calculated by dividing *D* by its maximum possible value, given the allele frequencies at the two loci (Lewontin and Kojima, 1960). As a result, complete LD, when $D' = 1$, occurs if, and only if, two SNPs have not been separated by recombination (or recurrent mutation or gene conversion) during the history of the sample. Values of $D' < 1$ indicate that ancestral LD has been disrupted. However, the relative magnitude of values of $D' < 1$ has no clear interpretation and therefore should be avoided as a measure of LD, especially *D'* values between 0.3 and 0.7 (Clark et al., 1998, Weiss and Clark, 2002).

LD has been widely studied in various domestic animal species (Khanyile et al., 2015, Kim and Kirkpatrick, 2009, Meadows et al., 2008, Muñoz et al., 2019, Prieur et al., 2017), and in many plant genomes (Andrade et al., 2019, Flint-Garcia et al., 2003, Remington et al., 2001). In nematodes, there have only been a few (very limited) analyses of LD. These include a comprehensive analysis of genetic diversity and the construction of a first filial generation (F1) genetic map in the parasitic helminth, *H. contortus*, carried out by Doyle et al. (2018). Another study by Cutter et al. (2006) revealed high nucleotide polymorphism and rapid decay of LD in wild populations of *C. remanei* (Cutter et al., 2006). Evans and Anderson (2020) used linkage mapping on recombinants derived from a cross between the laboratory strain (N2) and a wild strain (CB4856) to identify a single overlapping QTL on chromosome V that influences the responses of *C. elegans* to eight chemotherapeutic compounds. They discovered that the drug-response QTL overlapped with an expression QTL hotspot that contains the gene scb-1, previously implicated in bleomycin response in the parasite (Evans and Andersen, 2020). A limited study in *O. volvulus* focused on LD between short regions within and to either side of the

P-glycoprotein gene by (Ardelli et al., 2006). However, the extent of LD and haploblock structure has not been characterized across an entire chromosome of the parasitic nematode *O. volvulus*.

1.8.Genome imputation.

Sample size is a limiting factor in association studies (Spencer et al., 2009), but large increases in sample size to compensate for weak LD between a marker and a causative QTL can be impractical for many organisms. This is particularly true for studies that aim to use whole genome sequencing to identify QTLs associated with ivermectin response in *O. volvulus* because the limited funding and ethical concerns for surgical removal of worms from people in affected African countries, the difficulty of assessment of the definitive drug response related phenotype (embryogram counts of microfilariae in uteri), and the cost of genomic sequencing all limit the number of samples available for genome association studies (Doyle and Cotton, 2019, Hedtke et al., 2019).

In addition, whole genome sequencing or genotyping of all the identified variants in the genome is expensive and almost certainly impractical for most helminth parasites of humans, given the often-poor quality and low concentration of genomic DNA that can be prepared from worms isolated under field conditions in developing countries. However, as the genotypes at nearby markers are usually correlated (that is, they are in LD), it may be possible to scan the genome using a much smaller marker set with only a modest loss of power to detect selection while minimising the amount and quality of DNA that is required. More importantly, the power of relatively small (compared to the whole genome) number of markers to detect selection can be augmented significantly using methods that predict the genotype at un-typed/missing loci from a set of reference genotypes (known as the ‘reference panel’) by imputation (Nothnagel et al., 2009). To design such studies, it is necessary to have a detailed understanding of the structure and extent of LD across the genome, both to choose suitable reference and genotyping marker sets and to design appropriate methods of statistical analysis (Li et al., 2009).

GWAS using genomic imputation can be accurate and well calibrated if suitable reference and genotyping panels are available for the populations of interest. It has been used convincingly in many GWAS of complex human diseases (Marchini and Howie, 2010, Marchini et al., 2007, VanRaden et al., 2013). Examples include the HapMap project, where a reduced number of SNPs allowed the successful prediction of human leukocyte antigen (HLA) class I and II gene alleles (International HapMap 3 Consortium, 2010, International HapMap Consortium, 2003), and a HapMap CEU-based imputation study that reliably inferred missing genotypes in a population of northern European descent, even in variable regions such as the extended major histocompatibility complex (MHC) (Nothnagel et al., 2009). Another example is in a whole genome sequence study of the Iceland population, described above in section 1.5 (Gudbjartsson et al. 2015). Gudbjartsson et al. were able to impute variants into 104,220 individuals down to a minor allele frequency of 0.1%. In addition to a comprehensive understanding of the structure of the Icelandic population, this study demonstrated the use of imputation to discover associations between variants in sequence and phenotypes (Gudbjartsson et al., 2015). Imputation is also effective in livestock populations because of strong LD due to inbreeding. It is also usually performed within breed as linkage phase does not necessarily hold across more diverse populations (Nothnagel et al., 2009).

Imputation is new to filarial nematode population genetics studies, and reference panel datasets has not been developed that could be used to impute missing genotypes in these animals. Reference panels have been developed for other organisms where genotype-phenotype associations are more routinely applied, including in humans (the Human Genome Diversity Project (Cavalli-Sforza, 2005), the HapMap Consortium (International HapMap 3 Consortium, 2010), and the 1000 Genomes Project (1000G) (Sudmant et al., 2015)), Sheep - 5K, 50K, and HD panels (Ventura et al., 2016); and cattle (the 1000 Bull Genomes Project (Run 6.0) (Daetwyler et al., 2014), BovineSNP50 (SNP50) BeadChip (Illumina, San Diego, CA) (Matukumalli et al., 2009), and the BovineHD (Illumina, San Diego, CA) array (Wiggans et al., 2012)). However, genomic imputation has not been tested in GWAS of parasitic nematodes even though, it could be a powerful tool for minimizing costs associated with genetic-based screening for SOR in *O. volvulus* or for drug resistance in other helminths.

1.9. Summary, Hypothesis and Aims of this Thesis.

To conclude, global control and elimination of helminth diseases relies on the efficacy of anthelmintic macrocyclic lactone. Ivermectin is currently the only drug used by MDA programs to eliminate *O. volvulus* transmission in endemic foci. However, the long-term success of these MDA programs is limited by the emergence and spread of SOR of *O. volvulus* to ivermectin in some of these foci. Consequential to continuous drug pressure and the potential migration among endemic regions, additional parasite populations have the potential to develop SOR to the drug (or loss of efficacy or resistance). To face this growing threat, detailed knowledge of ivermectin response mechanisms is required to improve diagnostic tools for field identification of SOR and to subsequently update elimination strategies used by MDA programs. Resistance to ML is already a problem in controlling *D. immitis* (Bourguinat et al., 2015) infecting dogs and *T. circumcincta* (Choi et al., 2017) infecting sheep, and GWAS has been used successfully to identify multiple loci that are responsible for ML resistance in these parasitic nematodes. Similarly, QTLs that are under selection for SOR have been identified in *O. volvulus* (Doyle et al., 2017, Hedtke et al., 2017) using GWAS. That said, GWAS is complicated in *O. volvulus* because of its life cycles which does not allow for functional studies, limited sample size due to challenges with collecting high-quality parasite genomic DNA from humans in low-resource settings and limited existing molecular and genetic tools. To alleviate this, I leveraged the repository of *O. volvulus* single worm whole genome sequences generated by Choi et al. 2016 and Hedtke et al. 2017 to develop a methodological framework that could improve GWAS for ivermectin response in *O. volvulus.* This methodological framework tests amplicon resequencing to validate putative QTLs, identifies LD structure between QTLs and causative SNPs, and genotype imputation to increase the visibility of the causative SNPs or loci.

In this thesis, I tested the hypothesis that SOR to ivermectin in *O. volvulus* is genetically determined, such that LD will exist around ivermectin-response loci that are under selection, and that this will allow rational experimental design for GWAS of *O. volvulus*. I propose to test thesis hypothesis using three complementary approaches:

1) Determine the extent of genetic association between the ivermectin-response phenotype and the genotype at non-synonymous SNP loci within QTLs that have strong support from previous GWAS, using amplicon resequencing to increase the sample size (Chapter two).

2) Characterize patterns of LD decay and structure surrounding QTLs and more generally across *O. volvulus* autosomes, for the first time in *O. volvulus* (Chapter three).

3) Explore the feasibility and accuracy of imputation by making use of two different sets for reference panels and test the success of imputation in improving the power

of association of SNPs with ivermectin response, which has not been previously explored in filarial nematodes (Chapter four).

Chapter Five discusses the significance of each of the key results obtained from the experimental chapters and suggests future directions for research on the mechanisms of ivermectin response in *O. volvulus*.

Chapter Two

Amplicon-resequencing to validate ivermectin resistance candidate regions.

2.1. Introduction

Onchocerca volvulus, the cause of river blindness, shows variation in response to the drug used in mass drug administration, ivermectin. Association Study using Wright's F-statistics to quantify genetic differentiation between worms with different ivermectin response, identified associations between genotype and sub-optimal response to ivermectin (SOR) in adult female worms from Ghana and Cameroon and suggested that SOR is a polygenic quantitative trait (Doyle et al., 2017). However, this study used low sequence coverage pool-seq data of a limited number of worms which resulted in a stochastic variation in allele detection in the worms (Doyle et al., 2017). An additional association study was performed on genome sequences of individual worms, also using F_{ST} to identify regions with strong association with SOR phenotype in *O. volvulus* from Ghana (Hedtke et al., 2017), but this study was also limited in sample size to only 48 worms. Neither study performed additional validation of the identified regions using increased sample size. In part, this was because whole genome sequencing of single worms was limited by the insufficient quality of the DNA extractions available. Thus, my study aims to find options that could be useful in improving the sample size for GWAS in *O. volvulus*.

F_{ST} is a widely used descriptive statistic in population genetics that has been used to identify regions, or Quantitative Trait Loci (QTLs), of the genome that have been the target of selection. Its application ranges from disease association mapping to forensic science (Holsinger and Weir, 2009). QTL is a location in the genome identified as associated with a quantitative trait in a population (in this case, SOR) (Doyle and Cotton, 2019). QTLs are often associated with polygenic traits and identifying them is important because they inform the genetic architecture of the phenotype.

In this chapter, I explored an amplicon resequencing approach to improve sample size for verifying the QTLs proposed by Hedtke et al. (2017). Amplicon sequencing enables targeted analysis of genetic variation in a specific region and is less sensitive to the quality of the genomic DNA extractions. Strategic sequencing of QTLs that might contain multiple genes and downstream/upstream regions of those gene can allow fine mapping of regions, detecting a wider variety of polymorphism types (including insertions and deletions), characterising rare alleles, and characterising linkage disequilibrium (LD) (Goode, 2011). The primary reason for using amplicon resequencing in this study, however, is because the genomic DNA was too degraded for whole genome sequencing. Additional advantage to this approach compared to whole genome sequencing is the reduced sequencing costs and shorter turnaround time. I targeted 20 regions for PCR amplification and sequencing within the elevated F_{ST} regions (that is, regions with F_{ST} values >5 standard deviation above the mean) identified in Hedtke et al. (2017) that strongly differentiated SOR from good responder (GR) phenotypes.

In this chapter, I describe the first independent study to attempt to verify SOR-QTLs observed in a GWAS of ivermectin resistance in *O. volvulus* from Ghana. My initial intention was to sequence an entire QTL that showed strong association with SOR with long-range amplicons. Long-range PCR is a flexible, fast, efficient, and cost-effective choice for sequencing candidate genomic regions, especially when combined with next-generation sequencing (NGS) platforms (Jia et al., 2014). Long range PCR allows for amplification of genomic DNA lengths up to 22 kb which in most cases cannot be amplified using conventional PCR methods or reagents (Theophilus and Rapley, 2002). Sequencing of sufficiently long amplicons would also allow the prediction of phase and estimation of LD within each amplicon (Guo et al., 2006). Unfortunately, the same DNA quality that prevented effective whole genome sequencing also prevented amplification of long fragments from the gDNA. As a result, I used much shorter amplicons (< 500 bp) to target specific nonsynonymous mutations within those putative QTLs (that is, the elevated F_{ST} regions) and not the entire QTL, as well as several that were not predicted to be associated with SOR. This sequencing approach may validate candidate genetic variants that would contribute to the phenotypic difference used to categorise worms as GR or SOR.

2.2. Materials and Methods

2.2.1. Study Data

The DNA templates used in this study consist of 218 adult female *O. volvulus* collected from 14 communities and along five (5) rivers that form part of the Black Volta River basin and Volta Lake in Ghana in West Africa (Figure 2.1). Worms were phenotyped based on embryogram data examining presence/absence of stretched microfilariae in the uteri of worms extracted via nodulectomy and were categorized into good, medium, poor, or very poor (Osei-Atweneboana et al., 2011), which has been further binned into 'good responder' (GR; that is, absence of microfilariae in the uteri) or 'sub-optimal responders' (SOR; medium, poor, or very poor; that is, presence of stretched microfilariae in the uteri) in this study (Table 2.1). Whole genomic DNA (gDNA) extraction was done for individual worms from the non-reproductive tissue as reported previously by (Armoo et al., 2017). The gDNA were quantified using a Qubit Fluorometer (ThermoFisher Scientific) and standardized to a concentration of 1ng/ul prior to use in PCR.

2.2.2. Choice and Location of Target Loci

Regions were chosen based on loci identified as associated with SOR in Ghana from a preliminary GWAS (Hedtke et al., 2017), based on F_{ST} analyses of genome sequences from 48 worms. To expand the number of worms that could be used for F_{ST} analysis, I selected regions within elevated F_{ST} that strongly differentiated GR and SOR worms (Figure 2.2) for targeted sequencing using PCR, as shown. I chose to focus on regions where there were non-synonymous SNPs that might be functionally relevant. Initially, I targeted the 3 QTLs located on two major autosomes, chromosome OM1 and OM4. I designed three primer pairs of approximately 8kb each for those (Table 2.2) and they are referred to as long-amplicons in this study.

Because difficulties were encountered when amplifying approximately 8kb from gDNA extracts of variable quality, short amplicons of approximately 500 bp long were developed to target SNPs that were non-synonymous within the same QTL as well as others on the OM1 and OM4 chromosomes. Twenty-five (25) SNPs of interest from a list of missense mutations within elevated-F_{ST} peaks (or QTLs) were captured for amplicon resequencing (Figure 2.2). Primers were designed across OM4 chromosome (11 amplicons) and OM1 chromosome (9 amplicons) with no amplicon of more than 500 bp (OM1 chromosome is represented by two large contigs, OM1a and OM1b) (Figure 2.2).

The total combined length of the 20 amplicons was 7948 bp (approximately 0.008% of the genome; Figure 2.2).

PCRs were performed on 138 phenotyped adult female worms using 3 long amplicons and 218 using the 20 short amplicons. Finally, for the final pilot study, 71 worms from which all short (<500bp) amplicons were consistently amplified were sequenced on the Illumina MiSeq platform. Table 2.1 shows the study samples for amplicon resequencing categorised by their phenotypes, community of origin, and river basin.

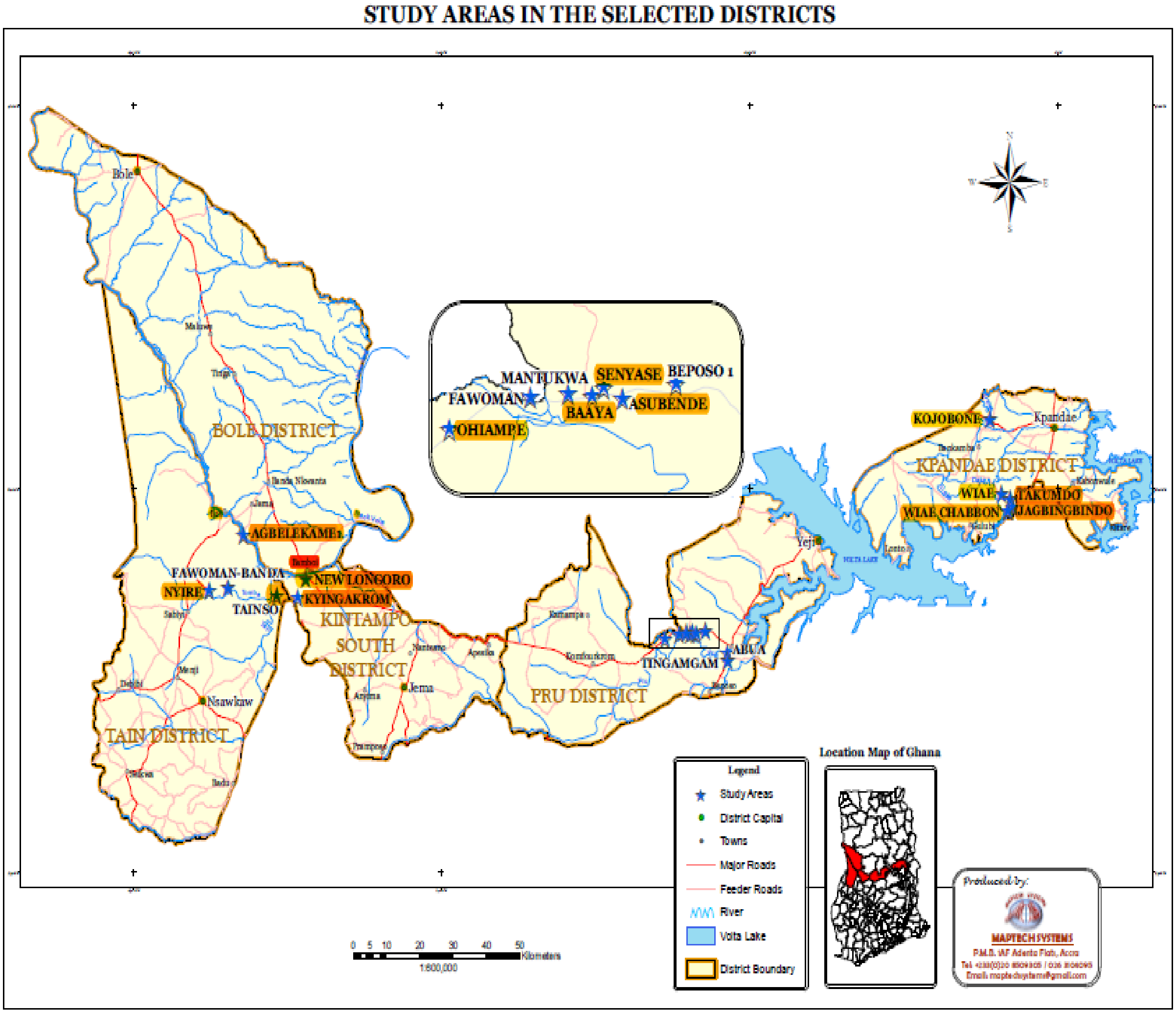


Figure 2.1. The map showing the locations of the sampling areas on the map of Ghana.

The highlighted areas are the communities where the *O. volvulus* were collected.

Table 2.1: Study sample for pilot amplicon resequencing experiment on Illumina MiSeq sequencer arranged by categories, code/ID and the number of worms in each category.

Categories	Code/ID	Number
Response Phenotype	Good responders (GR)	25
	Sub-optimal Responders (SOR)	46
	Total	**71**
Community	Agbelekame 1 (AB1)	5
	Agbelekame 2 (AB2)	11
	Asubende (ASU)	5
	Baaya (BAY)	2
	Jagbengbendo (JAG)	12
	Kojoboni (KOJ)	2
	Kyingakrom (KYG)	8
	New Longoro (NLG)	2
	Nyire (NYR)	6
	Ohiampe (OHP)	2
	Senyase (SEN)	1
	Takumdo (TAK)	8
	Wiae (WIA)	2
	Wiae Chabbon (CHA)	5
	Total	**71**
River basin	Black Volta (BV)	26
	Daka (DK)	21
	Pru (PRU)	10
	Tain (TA)	8
	Tombe (TM)	6
	Total	**71**

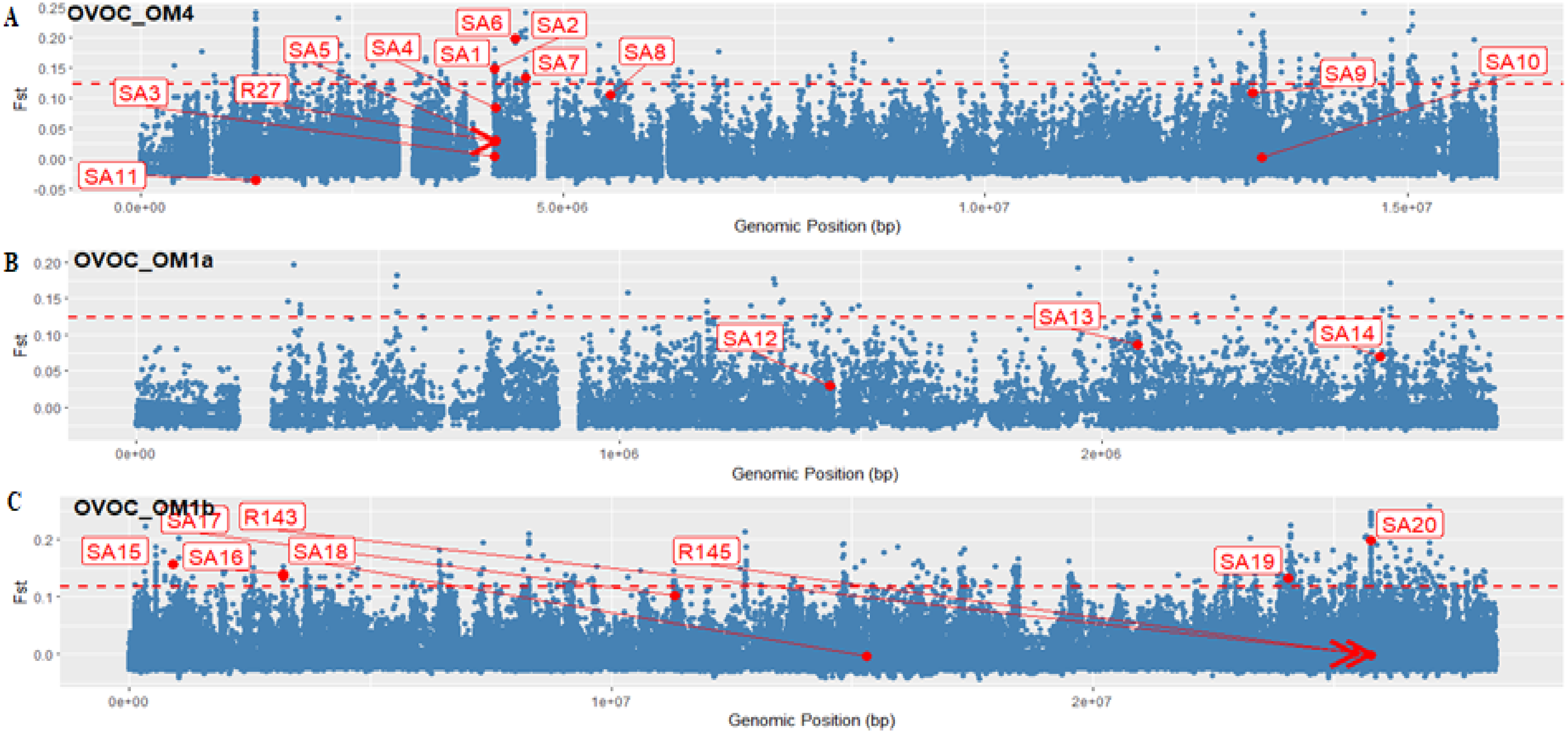


Figure 2.2. Physical location of amplicons on the nuclear genome.

Graph showing the degree of genetic differentiation (F_{ST}) between GR and SOR worms across the *O. volvulus* nuclear genome from previous GWAS by Hedtke et al. (2017). "Panel A": OVOC_OM4; "Panel B": OVOC_OM1a and "Panel C": OVOC_OM1b chromosome. The X-axis is the positions of SNPs across autosomes in the *O. volvulus* nuclear genome. The Y-axis is the GR vs SOR Weir F_{ST} values. Blue points are the position of SNPs with corresponding F_{ST} values. The red horizontal line is five-standard deviation above the mean F_{ST}. This serves as a threshold. Points above the line show genomic positions having strong differentiation between GR and SOR. The red dots and arrows point to the location of target SNPs. “SA” are amplicons of approximately 500bp while “R” are amplicons of approximately 8kb.

Table 2.2: Long (approximately 8kb) and Short (approximately 500bp) Amplicon ID, description (long or short), chromosome, genomic coordinates, and size.

Amplicon ID	Description	Chromosome: Genomic coordinates	Amplicon size (bp)
R143	Long	OVOC_OM1b: 25417588 - 25423102	5514
R145	Long	OVOC_OM1b: 25753231 - 25761066	7835
R27	Long	OVOC_OM4: 4189206 - 4197154	7948
SA1	Short	OVOC_OM4: 4191754 - 4192133	380
SA2	Short	OVOC_OM4: 4192145 - 4192597	476
SA3	Short	OVOC_OM4: 4194013 - 4194465	453
SA4	Short	OVOC_OM4: 4196327 - 4196666	340
SA5	Short	OVOC_OM4: 4199230 - 4199498	268
SA6	Short	OVOC_OM4: 4433648 - 4433886	239
SA7	Short	OVOC_OM4: 4551496 - 4551744	248
SA8	Short	OVOC_OM4: 5561435 - 5561781	347
SA9	Short	OVOC_OM4: 13168575 - 13168898	324
SA10	Short	OVOC_OM4: 13277211 - 13277600	390
SA11	Short	OVOC_OM4: 1358710 - 1358979	270
SA12	Short	OVOC_OM1a: 1436794 - 1437052	259
SA13	Short	OVOC_OM1a: 2073406 - 2073762	357
SA14	Short	OVOC_OM1a: 2577110 - 2577287	178
SA15	Short	OVOC_OM1b: 907915 - 908231	317
SA16	Short	OVOC_OM1b: 3194010 - 3194410	401
SA17	Short	OVOC_OM1b: 11320747 - 11321065	319
SA18	Short	OVOC_OM1b: 15299740 - 15300014	275
SA19	Short	OVOC_OM1b: 24029442 - 24029781	340
SA20	Short	OVOC_OM1b: 25758816 - 25759105	290

NB: “R” are amplicons of approximately 8kb.

“SA” are amplicons of approximately 500bp

2.2.3. Primer design and dilution.

Primers were generated using National Centre for Biotechnology Information (NCBI) PrimerBlast (Altschul et al., 1990) and CLC Genomics Workbench 20.0 (https://digitalinsights.qiagen.com). The qualities of the primers were confirmed using the OligoAnalyzer tool by Integrated DNA Technologies (IDT) (https://sg.idtdna.com/). Characteristics used for selecting primers are detailed in Table 2.3. Designed primer sequences were compared to existing whole genome sequences for the presence of population genetic variation using CLC Genomics Workbench. Primers that contained an identified SNP were either excluded, or a degenerate primer used, or were re-designed to exclude the SNP.

2.2.4. PCR Optimization and Amplification.

2.2.4.1. Primer Optimization of PCR reaction conditions.

PCR reaction annealing temperature was optimized to reduce primer-dimer formation and to increase the efficiency and specificity of the amplification process. A negative control (no template control) was used to detect reagent contamination or background signal. Positive controls (successful amplicons of the same size) were also used as quality control check. All amplifications were carried out in a gradient PCR machine (TaKaRa PCR Thermal Cycler; ThermoFisher Scientific, Waltham, Massachusetts, United States).

2.2.4.2. Amplification of the long (approximately 8kb) amplicons.

GoTaq® DNA polymerase was used to amplify the long amplicons using the manufacturer's standard protocol. PCR reactions were performed in a final volume of 20 µl containing 2 ng of genomic DNA, 10 µl of GoTaq® Long PCR Master Mix (2X) (Promega, US), and 0.1 µM of each primer. The protocol comprises of an initial activation step of 95°C for 2 minutes, 30-35 cycles of 94ºC denaturation for 30 s, annealing temperature (52ºC, 58.6ºC, and 56ºC for amplicons R145, R143, and R27, respectively (Table 2.4)) for 30 s, and 70°C extension for 1 min / kb, with a final extension step at 72°C for 10 min. Samples were then chilled to 10°C.

Table 2.3. Characteristics used for selecting primers.

Primer Length	18-20 bp
Annealing Temperature (Tm)	52 to 58 °C
GC content	40-60%
GC clamp	< 3 G's or C's
Base-pair repeats	no more than 4
Tm between primers	< 5 °C
3' end hairpin	ΔG ≤ -2 kcal/mol
Internal self-dimer	ΔG ≤ -6 kcal/mol
3' end cross dimer	ΔG ≤ -5 kcal/mol
Internal cross dimer	ΔG ≤ -6 kcal/mol

2.2.4.3. Amplification of the Short (approximately 500bp) amplicons.

PCR reactions were set up for the short amplicons as per the IMMOLASE PCR (Bioline Reagents Ltd) protocol with 0.25 µM of each primer, 0.4mM of dNTPs, 3mM of $MgCl_2$, 10X buffer, and 1U IMMOLASE enzyme in a final 20 µl volume reaction containing 2 ng of genomic DNA. PCR consisted of an initial denaturation step at 95°C for 10 minutes, 35 cycles of denaturation at 95 °C for 30 seconds, annealing at a temperature range of 52 to 61.2ºC for 30 seconds depending on the amplicon (Table 2.4), and extension at 72ºC for 30 seconds / kb. A final extension step was performed at 72°C for 5 minutes and samples were chilled to 10ºC.

Table 2.4. Long and short amplicon primer sequences and optimal annealing temperature.

Amplicon ID	Forward primer Sequence	Reverse primer Sequence	Optimal Tm
R145	5 '- GGACCAGCTTTGTTGGCTTC - 3'	5' - CCGTTGAAACACGACCAGGA - 3'	52 C
R143	5' - GTGAGTTTCTTCCTTCTGCTGC - 3'	5' - ACGACAACACGTCAAACCAC - 3'	58.8 C
R27	5' - TATTTCTGGACTGGTTGG - 3'	5' - CATGATTTTGGATTCGTTGG - 3'	56 C
SA1	5' - AGGTGCACGTCATTCAGTGT - 3'	5' - GGAAAGACGGGAATATYGACCA - 3'	58.8 C
SA2	5' - TGCGTAAAGCACTCAGGTGAT - 3'	5' - CCCTATTTTARCGGATTTGCTAGG - 3'	58.8 C
SA3	5' - GTTGTGGGAAATATTGAGC - 3'	5' - TCAAAAACCCTCATGCCG - 3'	58.8 C
SA4	5' - ATCAAGATTGGTTCCGAAGA - 3'	5' - TGGCCCATTTCACCATTACG - 3'	54 C
SA5	5' - TTTCCTCCCTGATTATTTCTGC - 3'	5' - TTCAATCCAAAACAGTCCACC - 3'	52 C
SA6	5' - ACCAATACTCCATGCTTGTGC - 3'	5' - AGGAATGGTTATGGGMGGGAAT - 3'	58.8 C
SA7	5' - TTTTTAGCAGCGAGCGGGA - 3'	5' - AACCTAACCTCCATGAAATTCTGC - 3'	58.8 C
SA8	5' - TGAGCACAGTATCAGAAGAC - 3'	5' - GGTTGCTTGGATAAAACTGG - 3'	56 C
SA9	5' - CGTTTTCGGCAATTCATCTT - 3'	5' - CAGGCATCTTCCGTTTCTTT - 3'	54 C
SA10	5' - TGGTCATCCTAACGAAATGG - 3'	5' - AGAAACCAACCTGGCAATAA - 3'	54 C
SA11	5' - ACAGCCTTTAGAATTTTCCCMGG - 3'	5' - TTTTTGTTGCAGCTTTCGGC -3'	58.4 C
SA12	5' - TCAGCCAGCGAATTGAACTT - 3'	5' - ACTGCCTGCTAAAATGCGAG - 3'	60.2 C
SA13	5' - GTTCGAGAGCCGTCACAAAA - 3'	5' - TGTCTGAAGTGAGAAAACCTCG - 3'	60.2 C
SA14	5' - TTTTATGTACCGAAGCAAAGGC - 3'	5' - TGTHTGTCTGGAATTGAGCGT - 3'	58.4 C
SA15	5' - TCTTTCCATGAAATTATTGCTCAAA - 3'	5' - GTTGGTTTAACCGCAGCATT - 3'	58.8 C
SA16	5' - ACGTAACAAATCTCGCCTGGA - 3'	5' - CATTCGTTGCTTTGACCTGGA - 3'	58.8 C
SA17	5' - AGGCTCAAGTCGTATGGCAA - 3'	5' - TGCTTTGGTACTTCGTCGCA - 3'	61.2 C
SA18	5' - TCAGTCTGGCATTGGTATTGGA - 3'	5' - CGCTTGCATCAATCTATCCGT - 3'	61.2 C
SA19	5' - GCTGCCTTCTCCCGAGTAAA - 3'	5' - TAGGACTTGAATTGCCCGTCG - 3'	61.2 C
SA20	5' - CAACATTCCCCACAAAACC - 3'	5' - TGGATGTGATATGGAAAAGG - 3'	56.5 C

2.2.5. Confirmation, Quantification and Validation of Amplicons.

To verify the results of each PCR reaction and confirm yield, each product was visualised on a 0.8% and 1% agarose gel for large and short amplicons respectively, with 0.03 µL/ml of GelRed (Biotium, Hayward, California, United States) added for band visibility. After running the gel at 80V or 100V for large and small amplicons respectively, they were viewed under UV illumination with either a 1kb or 100 bp DNA ladder (New England Biolabs, Ipswich, Massachusetts, United States) to confirm product size.

For the purpose of Sanger sequencing, amplicons that gave clean and clear bands were excised from the gel and purified according to the Wizard R SV Gel and PCR Clean-up System (Promega, Madison, Wisconsin, United States) protocol, with an extended incubation period of 1 hour. After clean-up, the purified PCR products were quantified using the Qubit™ reagent on Qubit 3.0 Fluorometer (ThermoFisher Scientific, Waltham, Massachusetts, United States).

2.2.6. Amplicon sequencing.

2.2.6.1. Sanger Sequencing.

For confirmation purposes, long-amplicon (1) and short-amplicon (9) were selected for Sanger sequencing: 20 µl of each purified PCR product and 20 µl of 10 µM of the corresponding forward primer was sent to Macrogen (Korea) for sequencing.

2.2.6.2. Next Generation Sequencing – MISEQ.

Twenty short-amplicons were prepared for each of 71 worm samples such that a total number of 1420 PCR reactions were processed for sequencing using the Nextera® XT (Illumina, Inc., San-Diego, California, United States) library preparation kit, which is designed for sequencing amplicons, small genomes, and plasmids, and then sequenced using 300-bp paired-end chemistry on an Illumina MiSeq sequencer with v3 reagents and 0.01 flow cells. To enable my target of interest with short amplicons to be sequenced, optimized primers were made “NGS compatible” by adding the Illumina sequencing primer and flow-cell adapters according to an in-house protocol. Next, index sequences were added to the samples before multiplexing, using Nextera® XT (Illumina, Inc., San-Diego, California, United States) indices followed by pooling of amplicons with amplicons from experiments on other species so that they were each at approximately equal concentrations with a final concentration of 2 nM. Quantification was done using PicoGreen® (ThermoFisher Scientific, Massachusetts, United States) on the CLARIOstar® (BMG LABTECH, Offenburg, Germany) before and after pooling PCR products and to quantify the library before sequencing.

2.2.7. Data and Bioinformatics.

2.2.7.1. Sanger Sequences.

Visual inspection, trimming and categorizing of base calling quality of the amplicons from Sanger sequence was done on CLC Genomics Workbench 20.0 (https://digitalinsights.qiagen.com). Heterozygous sites were inferred by observing secondary peaks using the CLC Genomics Workbench and confirmed with the Poly Peak Parser software (Hill et al., 2014): a Sanger sequencing file (abif or scf) containing a region of homozygous peaks followed by double peaks and a reference sequence for the region were imported into the Poly Peak Parser software. The software's default peak ratio cut-off parameter for calling heterozygous vs. homozygous positions (0.33) was used without exception on all of the amplicons. A reference sequence (from the *O. volvulus* nuclear reference genome (version 3) (Choi et al., 2016)) for each amplicon was provided and trimmed to match the length of the chromatogram. The percentage identity with reference was recorded. All positions in the sequence were examined and any dual peaks higher than the threshold of 0.33 were called as heterozygous sites.

2.2.7.2. Illumina sequences.

Thousands of sequences per amplicon were generated from the MiSeq run. FASTQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk) was used to perform quality checks to ensure that the raw data did not indicate a significant number of sequencing errors or any sequence failures due to inadequate starting material.

De-multiplexed paired-end sequence reads were trimmed using Trimmomatic v.0.32 (Bolger et al., 2014) such that only reads with PHRED score ≥ 30 and minimum size of 150 bp were retained. Leading and trailing low quality or N bases (below quality 3) were removed and any base with average quality of <15 was cut off using a sliding window of 4-base pairs.

Trimmed reads were aligned to the selected amplicons on the *O. volvulus* nuclear reference genome (version 3) (Cotton et al., 2016), using the Burrows-Wheeler Aligner (Li and Durbin, 2009). A custom *O. volvulus* reference was generated by combining sequences of only the selected amplicons on the nuclear genome. This was to ensure that variants were called from reads that mapped only to the desired amplicons. Samtools v1.3.1 (Li, 2011) was used to sort the mapping results. Custom Perl scripts were developed to use the output from samtools to calculate average depth and coverage across each amplicon given minimum depth. Variant calling was done using FreeBayes v1.0.2 (Garrison and Marth, 2012). Only variant sites supported by a minimum alternate count of 5, a minimum variant frequency of 0.25, a minimum depth of 20, and a minimum quality score of 30 were called. *Vcftools* (Danecek et al., 2011)) was used to further filter sites for quality and to remove sites and individuals with low read depth (less than 20) in more than 50% of the targeted variant sites.

2.2.7.3. Identifying Selection.

Vcftools (Danecek et al., 2011) was used to calculate Weir-and-Cockerham's F_{ST} (Weir and Cockerham, 1984) between GR and SOR and between each community and river basin. Association tests were carried out using plink2 v1.90b3 with the default parameters (Purcell et al., 2007). The ratio of non-synonymous to synonymous mutations (Ka/Ks ratio) and Tajima's D tests were carried out in the chosen coding regions to scan for evidence of selection using DNA Sequence Polymorphism (DnaSP) v6.10.04. (Rozas et al., 2017).

As an essential part of the QC steps, the presence of population stratification was tested using a multidimensional scaling (MDS) approach between the phenotypes and across river basins. The aim was to reveal groups of individuals that are genetically more similar to each other than expected.

All figures and tables were produced using *R studio* (R Development Core Team, 2013) using the *ggplot2* package (Wickham, 2016) and Microsoft Excel (Microsoft Corporation, Redman, Washington, USA).

2.3. **Results**

2.3.1. Long-range amplification.

Three large amplicons (approximately 8kb) were tested and run on an agarose gel to assess amplification success. The count of successful and failed amplification is given in Table 2.5. With amplicon R154 (7835bp), only 20 out of 138 DNA samples tested yielded a product of the expected length, 39 DNA samples gave non-specific amplification (that is, multiple bands) and 79 samples resulted in smears or no amplification (Table 2.5). To test consistency of results, the experiment on R145 was repeated using only the 20 samples that gave yields in the first experiment, and 10 that had non-specific bands, using the same long-range PCR conditions. 12 out of 30 samples gave the desired specific bands. The remaining 18 either gave non-specific bands or smears from the gel results (Table 2.5). This experiment was again repeated a third time using the same primers and long-range PCR conditions, and 13 samples resulted in the desired bands, while 2 samples that gave specific yield in the second trial resulted in smears (Table 2.5). In all, no one single DNA extraction was consistently amplified with high specificity using long-amplicon primers.

In summary, long amplicons were not successfully amplified consistently across all samples (Table 2.5), despite sufficient template (2ng) being used. As samples were extracted several years previously and stored at 4C, this is presumably due to DNA degradation and fragmentation.

2.3.2. Short-range amplification.

Short amplicons (approximately 500bp) were developed to target non-synonymous SNPs within the regions of elevated genetic differentiation between GR and SOR with the aim of finding candidate causative SNPs. An initial consistency check was done by running four primer pairs on all 218 DNA samples. 171 DNA samples were amplified successfully with specific yield of the expected length with the four primers (Table 2.5).

For the pilot experiment described in this chapter, DNA extracts from 71 phenotyped worms from those 171 that amplified successfully were chosen for further assessment for all using all 20 primer sets. Figure 2.3 shows the gel results for one of the short amplicons (SA1 - 380 bp), which successfully amplified across the 71 samples prior to illumina indexing step.

Table 2.5. Consistency check on the primers.

Amplicon ID	Number of DNA samples tested	Number of samples that gave specific amplification	Number of samples that failed
LONG AMPLICONS (~ 8kb)			
R145	138	20	118
	30	12	18
	30	13	17
Shared		**0**	
SHORT AMPLICONS (~500 bp)			
SA1	218	198	20
SA3	218	193	25
SA4	218	191	27
SA5	218	180	38
Shared		**171**	

Consistency check of long-amplicon primers (R145) and four short-amplicon primers (SA1, SA3, SA4 and SA5) on 138 and 218 DNA samples, respectively.

The category "Shared" is the number of samples that amplified with specific yield consistently across all primer sets**.**

Number of samples that failed comprises those that gave no yield and those that gave non-specific amplification after confirmation by gel electrophoresis.

NB: Inconsistency in the samples amplified was observed with the long-amplicons primers.

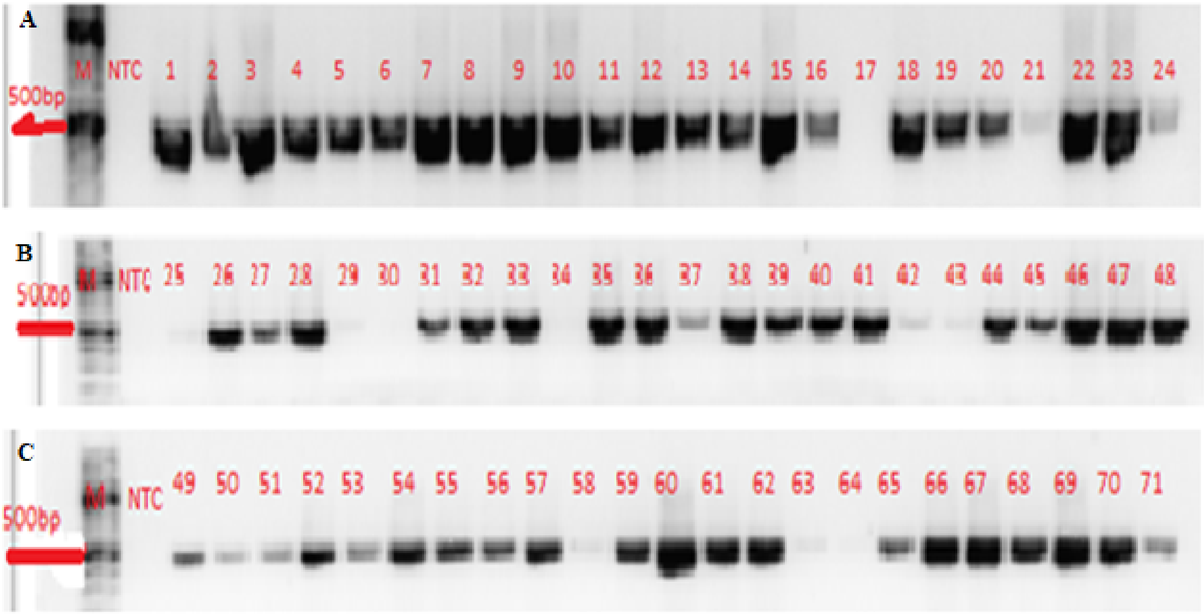


Figure 2.3: Agarose gel image of successful amplification of SA1 amplicon (380 bp) on 71 gDNA samples.

A: Lane 1: 100bp molecular marker, Lane 2: Non-Template Control (NTC) and Lane 3-26: DNA samples 1-24,

B: Lane 1: 100bp molecular marker, Lane 2: Non-Template Control (NTC) and Lane 3-26: DNA samples 25 - 48;

C: Lane 1: 100bp molecular marker, Lane 2: Non-Template Control (NTC) and Lane 3-26: DNA samples 49 – 71.

2.3.3. Assessment of Sanger Sequence Quality.

One long and nine short, purified PCR products were sequenced in one direction (with the forward primer) for 10 samples using Sanger sequencing to confirm the specificity of the primer design. CLC Genomic Workbench was used to assess the amplicons subjectively and to trim poor-quality bases at the 5’ and 3’ ends of the amplicons. The reference sequence for each amplicon was used by the Poly Peak Parser software for trimming to match the high-quality sequence’s chromatograms for each amplicon and the percentage identity with reference was recorded. Finally, heterozygous mutations were discovered by observing secondary peaks (that is, peak within peaks) using CLC Genomics Workbench. Table 2.6 describes the nucleotide sequence statistics of the output from Sanger after trimming, the number of bases with high quality, the percentage identity of those high-quality bases with the reference and heterozygous counts within the high-quality bases. After using CLC to trim poor-quality bases at the 5’ and 3’ ends of the amplicon R145, 820 bases out of R145 sequence output remained (expected amplicon length of 7835 bp), of which 718 (87.56%) bases were of high quality (that is, with quality score >30; they have 99.9% base call accuracy). R145 had 96.3% identity with its reference (Table 2.6).

The table also shows the report of the nine short amplicons sent for Sanger sequencing. After trimming for high quality, amplicons SA2 and SA6 had >93% of the returned trimmed sequences of high-quality bases (that is, bases with quality score >30 and they have 99.9% base call accuracy; trimmed length of SA2 and SA6 was 415 and 189 respectively). They have > 99% identity with the reference. Amplicons SA5, SA7 and SA8 failed because they had no (0%) identity with the reference, even though 29.28%, 10.75% and 3.37% of the trimmed returned sequences were of high quality in SA5, SA7 and SA8 correspondingly. Although amplicons SA1 SA3, SA4 and SA9 had greater than approximately 90% identity with the reference sequence, the percentage of high-quality bases after trimming were quite low (SA1 = 63.81%, SA3 = 44.33%, SA4 = 15%, SA9 = 76.59%) (Table 2.6).

The degree of mismatch in most of the amplicons is too high to be accounted for by polymorphism and suggests lack of specificity of the PCR. Specificity of the PCR reaction performed on some of the short amplicons was poor, while the PCR performed on the long amplicon had a percentage identity with the reference > 96.5%. Amplicons SA5, SA7 and SA8 had no (0%) identity with the reference (Table 2.6). Those short amplicons that had no percentage identity with the reference also failed to sequence adequately in the MiSeq run and were removed from downstream analysis.

The count and the percentage of the secondary peaks (that is, heterozygous mutations) within the high-quality bases in each amplicon are also recorded in Table 2.6. Heterozygous mutations were discovered by observing secondary peaks (that is, peak within peaks) using CLC Genomics Workbench and any secondary peaks higher than the threshold of 0.33 within the high-quality sequences were called as heterozygotes. The secondary peaks that met the 0.33 threshold and were thus called as heterozygotes ranged from 1 to 10 in the amplicons.

Table 2.6: Nucleotide sequence statistics: – description of each sequence output from Sanger after trimming, sequence with high quality, their percentage identity with the reference, and heterozygous counts.

Amplicon ID	Expected amplicon length (bp)	Sequence length after trimming (bp)	No (%) of bases with high quality after trimming	Identity of the high-quality sequences with reference (%)	Count of secondary peaks within high quality sequences (%)
R145	7835	820	718 (87.56%)	96.5	10 (0.32)
SA1	386	315	201 (63.81%)	93.0	10 (4.98)
SA2	476	415	407 (97.07%)	99.8	1 (0.25)
SA3	453	388	172 (44.33%)	92.6	9 (5.23)
SA4	340	260	39 (15%)	100	1 (2.56)
SA5	268	181	53 (29.28%)	0	6 (11.32)
SA6	239	189	177 (93.65%)	99.4	1 (0.55)
SA7	248	279	30 (10.75%)	0	3 (10.00)
SA8	347	192	6 (3.13%)	0	1 (16.67)
SA9	324	252	193 (76.59%)	100	2 (1.04)

NB: High quality sequence was calculated using the trimmed sequence with a cut-off quality score of 30.

2.3.4. Quality Check on the MiSeq data.

Modern high-throughput sequencers generate hundreds of millions of sequences in a single run. Before analysing these sequences to draw biological conclusions, simple quality control checks were performed to ensure that the raw data has low error rates and is not biased.

A total number of sequenced reads of 15,788,490 was processed for 20 amplicons of 71 samples. An average count of 222,373 reads were processed per sample. FastQC estimated a less than 0.2% error rate. Mean sequence quality (Phred score) was greater than 35 for all samples. N content is <5% for all samples indicating that the sequencer was able to make valid base calls with sufficient confidence. Mean length distribution across samples has its peak between 250-259.

After trimming and mapping, a total number of 4,168,624 bp was mapped for 20 amplicons of 71 samples with an average number of 58,713 bp were mapped per sample. Primer sequences was removed and the combined total for all samples and amplicons was 378,076bp (which is approximately 95% of the total expected consensus length without primer sequences). Variant calling was carried out and only individuals and amplicons with >50% of variants called wasretained. Eventually, based on the mean depth per amplicon across each sample as seen in Figure 2.4, mapping quality across all amplicons, and prior results from the Sanger sequencing, 3 amplicons (SA5, SA7 and SA8) were excluded from downstream analysis because they had low depth of coverage (<20) in 50 % of the samples. Four samples (GR_17, PR_30, PR_34 and VP_69) were excluded because they had low depth <20 in less than 50% of the remaining amplicons (Figure 2.4). Overall, 124 variants remained for downstream analysis across 67 worms and 17 amplicons (SA1, SA2, SA3, SA4, SA6, SA9, SA10, SA11, SA12, SA13, SA14, SA15, SA16, SA17, SA18, SA19 and SA20). 121 (97.6%) of these SNPs were biallelic, and 3 (2.4%) were multiallelic. The transition-transversion ratio was approximately 1.7. Fourteen (14) out of the 25 SNP loci that were the initial targets of this experiment were observed as polymorphic in this data. 50 (41.3%) of the 124 variable loci observed in this data were also polymorphic in the original GWAS, which means 74 new variants were called.

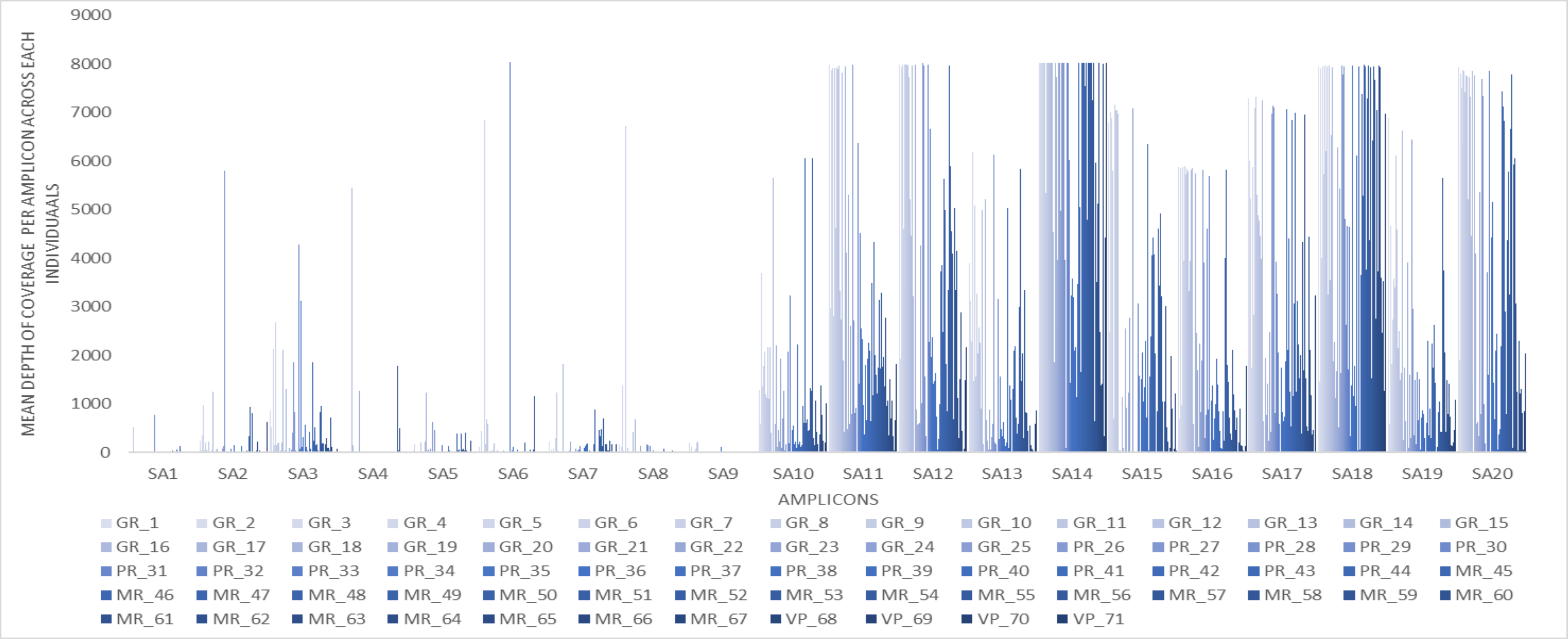


Figure 2.4. Mean Depth of each sample by short amplicon from the MiSeq experiment.

The x-axis is the individual short amplicons, and the y-axis is the count of depth for each sample by amplicon.

2.3.5. Genetic differentiation between ivermectin response phenotypes.

Analysis of samples by their community or river basin of origin did not reveal population genetic structure among the worms sequenced (among river basins: mean F_{ST} = 0.00148724; among communities: mean F_{ST} = -0.0131308) (Figure 2.5). Pairwise F_{ST} revealed low levels of differentiation between SOR and GR phenotypes (mean F_{ST} estimate between SOR and GR = 0.0041133) (Figure 2.6). Most of the SNPs were less than 2 standard deviations from the mean. Using a threshold of 5 standard deviations from the mean F_{ST} as a cut off (that is, z-score > 5.0, which is the threshold used in the previous GWAS; Hedtke et al. 2017), two outlier positions on amplicon SA9 were observed (Figure 2.6) with chromosomal positions OM4:13168647 and OM4:13168824 (the interval between them is 177 bp), but these were not significant after Bonferroni correction.

Functional analysis of the variants was done using SnpEff (Cingolani et al., 2012) to evaluate the possible functional effect of each SNP. Both high-F_{ST} SNPs are in the WBGene00248067 gene, which is predicted to have heme binding and peroxidase activity. OM4:13168647 is an intron variant with a potential modifier impact; while OM4:13168824 is both a missense and splice region variant.

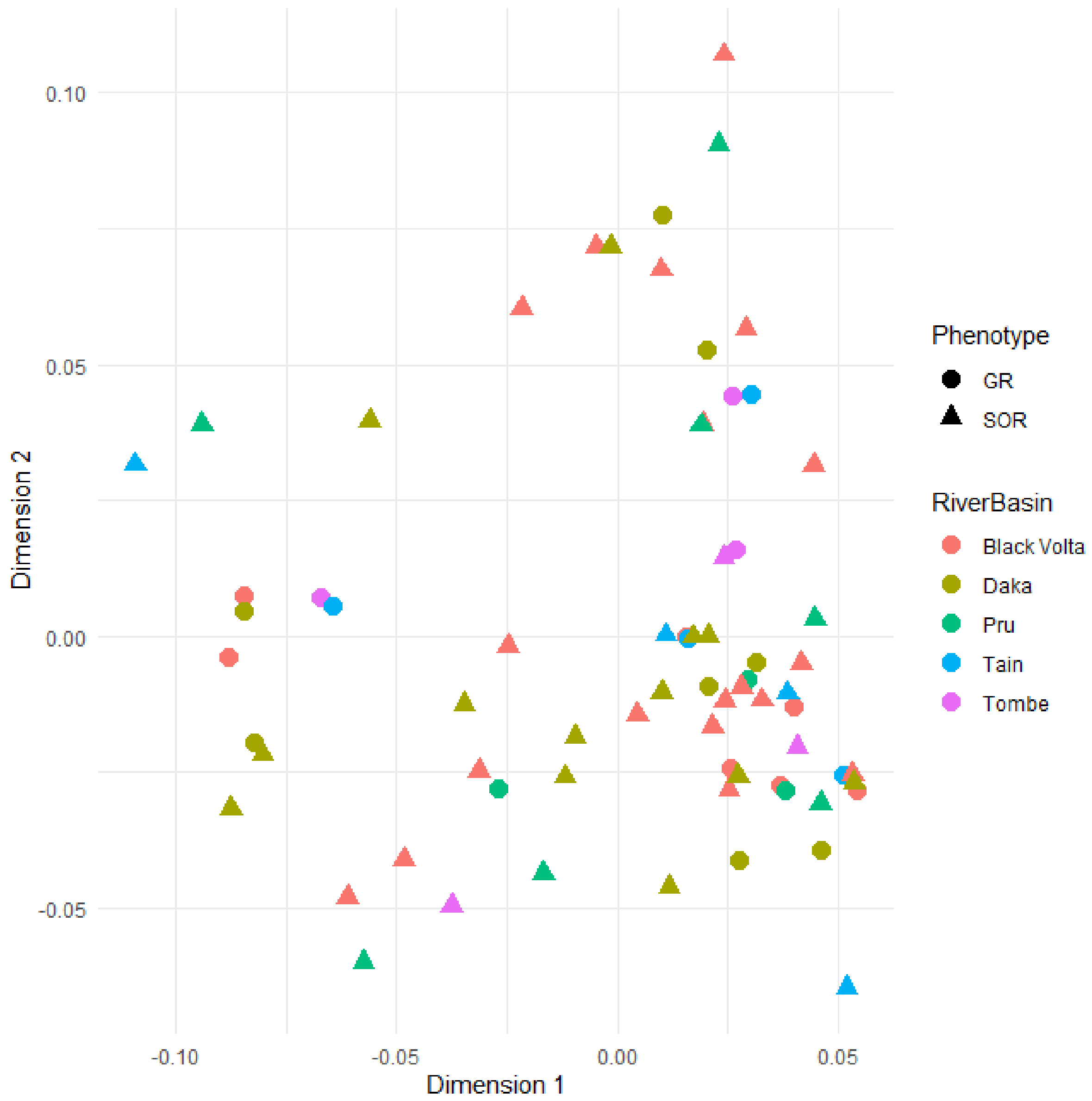


Figure 2.5. Non-metric Multidimensional Scaling plot across river basins – showing the geographical structure between phenotypes across river basins.

To identify population stratification (or relationship) between the phenotypes across river basins used in the association analysis.

No geographical nor population structure observed among the worms sequenced.

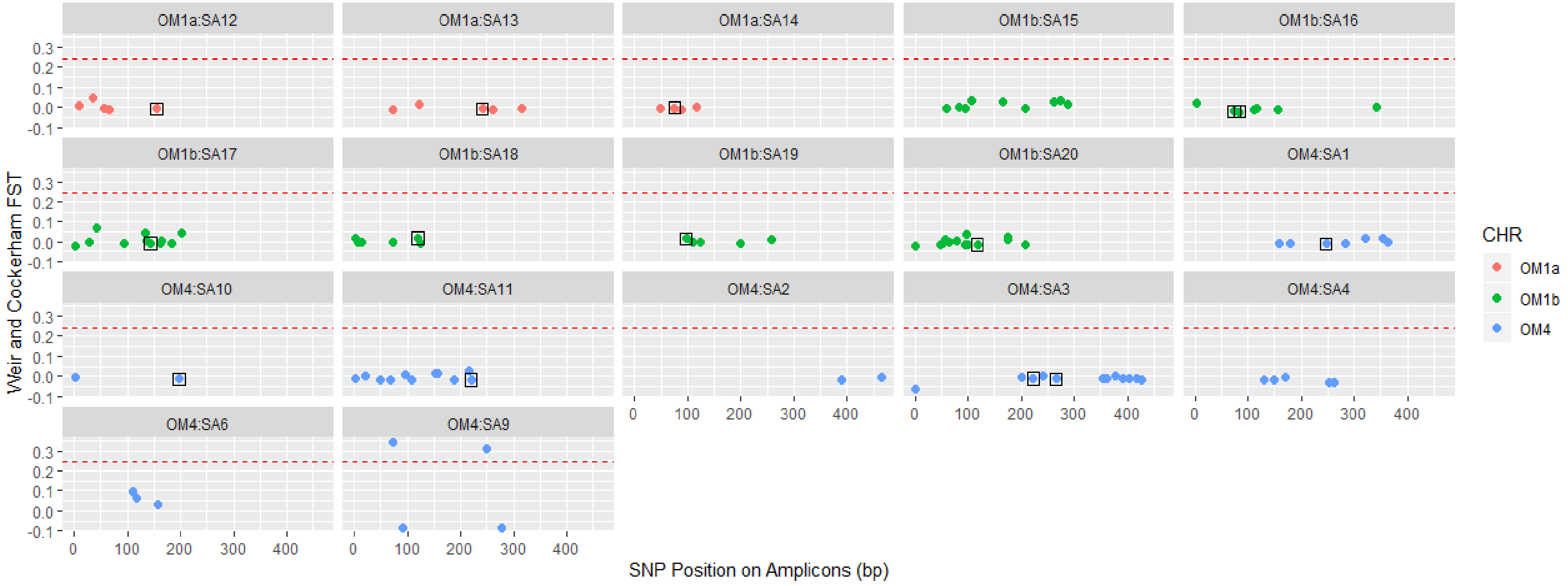


Figure 2.6. The genotypic differentiation between phenotypes (GR and SOR) across all mapped amplicons in my study.

F_{ST} values across all mapped short amplicons to validate variants with high levels of population differentiation between GR and SOR. The x-axes are the positions of SNPs across the short amplicons. The y-axes are the GR vs SOR Weir F_{ST} values for each short amplicon. The coloured points are F_{ST} values for SNPs within the amplicons coloured by chromosome (OM1a-Red; OM1b-Green; OM4-Blue) and the black rectangles highlights the F_{ST} values and positions of the initial target SNPs of interest within each amplicon. Outlier regions are above the red dashed line which is the average F_{ST} + 5 std threshold (F_{ST} = 0.25); they are found on amplicon SA9 at chromosomal positions OM4:13168647 and OM4:13168824 and are separated by approximately 200 bp.

NB: There are 17 amplicons, and they are distributed over 3 scaffolds (OM1a, OM1b and OM4).

2.3.6. Association Analysis.

Association analysis between genetic variation and SOR was carried out on the variants across all amplicons using the open‑source whole‑genome association analysis toolset *Plink2* v1.09 (Purcell et al., 2007). Only biallelic variants, (n = 121) with a total genotyping rate of 0.842376 (84%) were used. The genomic inflation est. lambda (based on median Chi-square) was 1.34593, signifying an increased false positive rate, as the value is greater than 1.00. Figure 2.7 is a Manhattan plot showing the SNPs and their -log10 P-value across amplicons. The figure shows that at a significance threshold of $P < 1 X 10^{-3}$, 2 SNPs showed association with SOR, but not significant after Bonferroni correction. Although the SNPs was different to those identified using F_{ST} (Figure 2.7) they are also on OM4 chromosome and the two high F_{ST} SNPs in Figure 2.7 are at approximately 9Mb distance away from these SNPs identified by *plink*.

The first is on amplicon SA5 at chromosomal position OM4:4199425 and the second is on amplicon SA6 at chromosomal position OM4:4433757; they are more than 200 kb apart. OM4:4199425 is 71 bp away from the initial target SNP on amplicon SA5 (chromosomal position OM4:4199354), showing that they may be linked with each other. OM4:4199425 is a variant that has a possible modifier impact and causes synonymous mutation on WBGene00246934 gene (which belongs to G-protein-coupled receptors (GPCR), rhodopsin-like superfamily). OM4:4433757 is 50 bp from the initial target on SA6 (chromosomal position OM4:4433807), also indicating that they may be linked with each other. OM4:4433757 is a variant in the WBGene00246959 gene, *Ovo-agl-1,* which is an ortholog of *C. elegans agl-1*, a glycogen debranching enzyme.

The lack of strong association of the initial missense variants targeted within the chosen coding regions with SOR indicates a need to further increase sample size to confirm the effect of the SNPs. I further increased the sample size by merging the worms from the GWAS with worms sequenced by this amplicon re-sequencing approach and the power of association of the SNPs at those loci was not improved either (results not shown). This is most likely that the QTLs from the discovery GWAS could be false positive because the analysis could not be replicated. The statistical power to detect an association largely relies on the effect size of the phenotype and number of samples used (As sample sizes increases, the phenotypic and genotypic variation in the sample cohort is also likely to increase). Nothing can be done to improve sample size in this context, and hence, the need to rely on large effect size of drug resistance phenotype because loci with small effect sizes may be difficult to characterize. However, precise phenotyping of individual *O. volvulus* parasite is difficult: uncertainties in the phenotypes collected from embryogram does not give high confidence in the ivermectin response phenotypes.

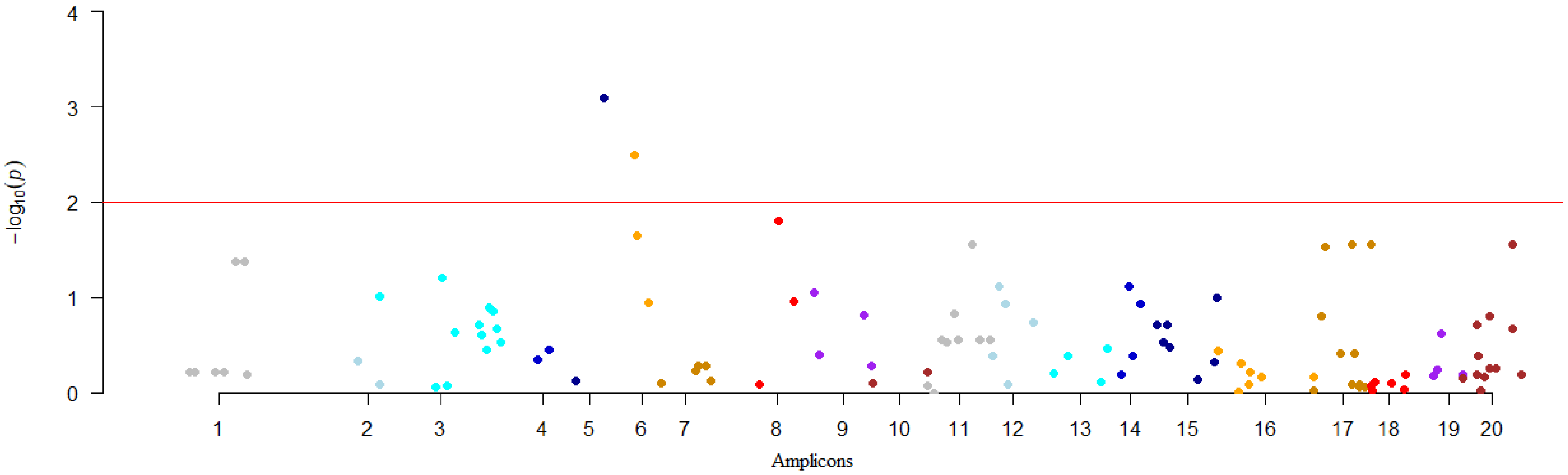


Figure 2.7. Manhattan plot depicting several strongly associated risk loci.

The X-axis shows the amplicons; the Y-axis shows the association level; and each colour coded dot represents a SNP and its location within an amplicon. Amplicons 1-11 are on chromosome OM4 and amplicons 12-20 are on chromosome OM1. The red horizontal line shows the p-value cut-off threshold. The top two SNPs above the red line are on amplicons SA5 and SA6, with chromosomal positions OM4:4199425 and OM4:4433757, respectively.

2.3.7. Test of neutral evolution in the selected nuclear genome regions of *O. volvulus*.

Neutrality tests were carried out on the chosen amplicons of *O. volvulus* to further assess deviations from neutral evolution at those amplicons. Table 2.7 shows the estimates of Tajima's D, estimates of Ka (the number of nonsynonymous mutations per nonsynonymous site), and Ks (the number of synonymous mutations per synonymous site) for the target amplicons in *O. volvulus*. The neutrality tests show deviations from neutral evolution at some amplicons. D=0 means no evidence of selection, D>0 means bottleneck (or balancing selection) and D<0 implies directional selection (or population expansion). Two amplicons that did deviate, and in opposite directions on OM4 and OM1b chromosomes, respectively, may reflect selection processes operating on those genes (Tajima’s D value for amplicon OM4_SA1 = -2.06 and for amplicon OM1b_SA16 = 2.57) (Table 2.7). This is not surprising, because, with respect to Tajima’s D, one could imagine either positive or negative values in connection with ivermectin response. Another confounder for these values could be because Tajima's D is also highly sensitive to demography. Similarly, variation in amplicon coverage could be underpinning this variation.

Likewise, the ratio of the proportion of nonsynonymous variants (Ka) to synonymous variants (Ks) can be used to infer the direction and the magnitude of natural selection acting on protein coding genes. A ratio greater than 1 implies positive (Darwinian) or directional selection (that is, driving change) while less than 1 is balancing selection and a ratio of 0 indicates neutral or no selection. From Table 2.7, the Ka/Ks ratio varies across amplicons, but none were statistically significant.

Table 2.7: Test of null hypothesis of neutral evolution (*Tajima's D*), estimates of Ka (the number of nonsynonymous substitutions per nonsynonymous site), and Ks (the number of synonymous substitutions per synonymous site) for the target amplicons in *O. volvulus*.

Amplicon ID	Population size	S	Tajima's D	Ka	Ks	Ka/Ks
SA1	67	6	-2.063662*	0.000812076	0.001140299	0.712160876
SA2	67	4	-0.309335	0	0.002016554	0
SA4	67	5	-0.947314	0.001502035	0.002431931	0.617630649
SA5	67	2	-1.430182	0.000531389	0	0
SA6	67	5	0.312532	0.006318996	0	0
SA7	67	8	1.796223	0.012183537	0.009958842	1.223388891
SA8	67	6	0.76347	0.003503573	0.0067891	0.516058545
SA9	67	5	-0.163366	0	0.004640118	0
SA10	67	5	-0.05081	0.002551967	0.002771235	0.920877399
SA11	67	8	0.765124	0.011608412	0.006955088	1.669053239
SA12	67	4	0.539015	0.002477612	0.005586793	0.443476571
SA13	67	4	0.711912	0.006633786	0.000824966	8.041282895
SA14	67	4	0.15403	0.006882225	0	0
SA16	67	5	2.573541*	0.008445907	0.004570828	1.847785001
SA17	67	12	0.234714	0.009781637	0.007391633	1.323339187
SA18	67	3	-0.197147	0.004691	0	0
SA19	67	5	0.887164	0.006858254	0.002907191	2.35906531
SA20	67	11	-0.494014	0.004834238	0.011942379	0.404796891

NB: S, number of segregating sites; Tajima’s D neutrality test; *Significant at $P < 0.05$; Ka, proportion of non-synonymous mutations; Ks, Synonymous mutations

2.4. Discussion

I described the findings from the first study to attempt the verification of SOR-QTLs observed in a previous GWAS of ivermectin resistance in *O. volvulus* from Ghana using an expanded sample size. I reported the findings from an amplicon re-sequencing pilot study that targeted 25 non-synonymous SNPs that were distributed over 14 QTLs defined in the previous GWAS (Hedtke et al. 2017).

2.4.1. Targeting long amplicons for validation studies.

Ideally long-range PCR would be used to confirm the association between F_{ST} and response to ivermectin, but this failed. Long-range PCR has been observed to speed up and simplify PCR for genomic mapping and sequencing and has facilitated studies in molecular genetics (Cheng et al., 1994, de Sousa Dias et al., 2013). When combined with sequencing, long-range PCR can provide a faster and more cost-effective tool for detecting genetic variations than whole genome sequencing (Knierim et al., 2011, Tan et al., 2012). In my study, attempts to consistently amplify long amplicons with sizes of 5.5 kb, 7.8 kb and 8.0 kb across several worms were unsuccessful, independent of the enzyme used for amplification. The difficulty in successfully amplifying long amplicons is highly likely to be associated with degradation of the gDNA samples: they had been extracted in year 2015 and stored at 4ºC. It is important to note that the question of gDNA storage is difficult - although storage at -20ºC is advised for long term storage the risk is that repeated freeze-thaw cycles will degrade the DNA faster than storage at 4ºC under most circumstances. So, it is a trade-off, particularly when the samples are being used for a variety of analyses on a regular basis (as these were). Another possible explanation is that the DNA extracts also include host (human) DNA, although estimates from whole genome data based on other extractions performed at the same time did not find extensive host contamination. It was also noted that only a small number of these samples were suitable for whole genome sequencing even when they were fresh and that the difficulty with long range PCR most likely reflects this initial quality problem: that is, if the sample failed for next generation sequencing library preparation then it will likely fail for long range PCR, Consequently, I have used smaller amplicons.

2.4.2. Discrepancies between the results from the previous genome-wide differentiation scan using F_{ST} using whole genome sequencing and the amplicon approach.

GWAS is a better approach when studying the genetic architecture of anthelmintic resistance in parasites than candidate gene approaches (that is, genes chosen, based on specific hypotheses concerning mechanisms of resistance to anthelmintic compounds) (Doyle and Cotton, 2019). Initial studies on the architecture of ivermectin response in *O. volvulus* took a candidate gene approach (Ardelli et al., 2006, Ardelli et al., 2005, Bourguinat et al., 2008, Nana-Djeunga et al., 2012, Osei-Atweneboana et al., 2012). However, genome-wide differentiation scan of ivermectin response in *O. volvulus* by Doyle et al. (2017) provided insight into the genomics of ivermectin response and population structure of the parasite and established that ivermectin response is a polygenic quantitative trait in which similar molecular pathways influence the extent of ivermectin response in the various parasite populations, and not discrete genes as proposed in candidate gene studies. Variants that differentiated GR and SOR parasites were found in several QTLs: however, there is the need for additional studies, including examining single whole genome sequences, to validate those QTLs rather than the pooled sequences used by Doyle et al. Thus, a second genome-wide differentiation scan using sequencing of single worms using samples from the same study sites was initiated (Hedtke et al., 2017).

The work described here aimed to verify the SOR-QTLs observed in the second GWAS of ivermectin resistance in *O. volvulus* from Ghana (Hedtke et al., 2017) using an amplicon re-sequencing approach on 25 non-synonymous SNPs that fell within the combined length approximately 7kb selected from 14 QTLs with an expanded sample size (n = 71). Amplicon resequencing that targets strategic QTLs is useful for fine mapping of regions and detecting a variety of polymorphism types (including insertions and deletions). More polymorphic sites were detected after sequencing the amplicons (74 new polymorphisms were detected). In my study, initially I picked an average of approximately 2 non-synonymous SNPs per amplicon from within each of the 14 QTLs selected (total = 25), but from the amplicon analysis, 14 out of the initial 25 SNPs were observed after amplicon sequencing in the following amplicons (SA1, SA3, SA10, SA11, SA12, SA13, SA14, SA16, SA17, SA18, SA19, SA20), and they show low levels of genetic differentiation between GR and SOR, that is, they all have low F_{ST} values. Even though the chosen non-synonymous SNPs were not directly associated with SOR after the analysis on the amplicon data, 2 new SNPs (OVOC_OM4:13168647 and OVOC_OM4:13168824) emerged from the new F_{ST} analysis on amplicon SA9 showing strong association with SOR. The physical relationship of these SNPs on amplicon SA9 to the nearest initial target SNP on amplicon SA10 is approximately 108 kb. Some non-synonymous SNPs that were found to be significantly associated with SOR in previous GWAS lost their association power after amplicon resequencing on a larger sample size. Precisely, there were 5 of the initial 25 that had high F_{ST} score (>0.13 or > 5 std) in the GWAS; and they were located on amplicons SA1, SA7, two on SA16, SA19 and SA20. But in the amplicon analysis, their F_{ST} scores were <0.02. This may be because (a) they may have been closely linked with a causative polymorphism or (b) they were highly differentiated between SOR and GR by F_{ST}, due to random chance associations that emerge when sampling sizes are not reflective of the population. This is particularly true for parasitic helminths which appear to have very large effective population sizes with correspondingly high levels of genetic diversity. Distinguishing between chance associations and true signal under these circumstances is the main reason for increasing sample size. These are common failures when using genome scans of F_{ST} to detect selection as reported by Holsinger and Weir (2009). Haasl and Payseur (2016) also reported that even though GWAS eliminates the bias often inherent in a candidate gene study, some selective events are difficult to identify by GWAS. For example, soft sweeps, polygenic selection, and selection targeting genetic variants such as microsatellites or copy number variants are more challenging to detect (Haasl et al., 2014, Innan and Kim, 2004, Pritchard and Di Rienzo, 2010). The primary reason that soft sweeps and polygenic selection are difficult to detect is that such sweeps leave little linkage disequilibrium (LD) around individual polymorphisms that contribute to the phenotype because the selection on any individual contributing polymorphism is weak. That explanation immediately raises the question of how much LD is there surrounding a single contributing polymorphism and led me directly to develop the chapter that followed this – which is the study on the structure of LD decay and haplotype blocks in *O. volvulus* chromosomes. Likewise, distinguishing the effect of demography and natural history from selection is challenging (Haasl and Payseur 2016). When selection targets complex phenotypes (when phenotypic variation reflects the action of multiple genetic variants – a typical situation in

O. volvulus), GWAS is less likely to succeed (Pritchard and Di Rienzo, 2010) and large sample size is extremely important.

In the case of soft selection acting on standing genetic variation of a polygenic or quantitative trait, the effect size of any single association is small, and this was observed in the original GWAS (Hedtke et al., 2017). A confirmation experiment with larger sample size would be needed to have the required power to confirm the effect. An attempt at increasing my study sample size by merging the worms from the GWAS with worms sequenced by this amplicon re-sequencing approach did not improve the power of association of the SNPs at those loci (results not shown). For complex traits such as ivermectin response in *O. volvulus* a GWAS based on whole genome sequencing generally identifies many polymorphisms, suggesting that individual polymorphisms each have a small effect (Goddard and Hayes, 2009). It could also be because, the GWAS (based on whole genome sequence) and this experiment (the amplicon resequencing project) used worms from two independent samples drawn from a single population, therefore, the association observed in the GWAS sample was a statistical artefact of small sample size that disappeared when sample size was increased. Also, the false discovery rate is often high, and so many significant associations are expected by chance when so many SNPs are tested. Goddard and Hayes corroborate this in their GWAS, where they found that SNP associations are most likely to be confirmed when the original GWAS used many animals (for example, >1,000) that were widely sampled from one breed, and confirmation of highly significant SNPs was carried out in an even larger sample of the same breed (Goddard and Hayes, 2009). Another factor for discrepancies between the GWAS on the whole genome and amplicon resequencing approach could be the effect of the binary categorization of the phenotypes into GR and SOR (Churcher et al., 2009). This categorization does not take into consideration the presence and the density of microfilariae in the skin. Bottomley et al. (Bottomley et al., 2016) found out that this method for phenotype classification is less sensitive for the determining the presence and extent of microfilarial density in the skin.

2.4.3. Identifying selection.

In this section, I have estimated four parameters that are commonly used to determine whether selection may have occurred: F_{ST}, Ka/Ks, Tajima's D, and association testing of individual SNP loci with phenotype.

The first criterion (elevated F_{ST}) - F_{ST} is useful for identifying which mutations or regions are candidates, but then further research needs to be done. F_{ST} is more than a descriptive statistic and a measure of genetic differentiation. It is directly related to the variance in allele frequency among populations and, conversely, to the degree of genetic resemblance among individuals within populations (Holsinger and Weir, 2009). A small F_{ST} means that the allele frequencies within each population are the same; conversely, a large F_{ST} means that the allele frequencies are different. If natural selection favours one allele over others at a particular locus in one population but not a second, the F_{ST} at that locus tend to be larger than at loci for which among-population differences are largely because of genetic drift. Genetic drift cannot, therefore, be ignored and the frequency with which fixation/loss happens across many loci will depend on the size of the population: two small populations can have a high degree of differentiation in allele frequencies with fixation occurring between them extremely rapidly. Furthermore, even in a large population there is still random loss or fixation and/or high F_{ST} due to genetic drift irrespective of selection. Thus, high F_{ST} is not, sufficient to detect selection. Selection differs fundamentally from drift because it occurs because of a genetic variant on the phenotype.

Therefore, in this study, I focussed on missense mutations which might be causative. If differentiation is due to selection, then a causative mutation will increase in frequency. So, the low F_{ST} observed across most SNPs examined is indicative that the initially selected amplicons had elevated F_{ST} due to chance and/or that the missense mutations are not causative. Figure 2.6 shows that 19 of the study amplicons have no evidence for selection using elevated F_{ST}. However, the SA9 amplicon (within chromosome OM4) showed some evidence of selection, with 2 SNPs within that amplicon having very high F_{ST} values (OVOC_OM4:13168647 and OVOC_OM4:13168824 with $F_{ST} > 0.13$). The two elevated FST SNPs are in the same amplicon. So, given the later results for LD, one would expect that they will be in LD irrespective of selection (they are, by definition, <500bp apart) and one expects that they are both elevated. What is unusual is that nearby SNPs in the same amplicon, that are closely physically linked, do not show similar evidence for selection, suggesting that random associations may still occur with this larger sample size. This could be due to random chance or could be because *O. volvulus* is sufficiently variable that there is selection on those variants regardless of genomic background (in other words, soft selection and low LD surrounding contributing sites).

Evidence to date suggests that ivermectin resistance in *O. volvulus* has been driven by soft selection on QTLs; that is, on polygenic and/or complex phenotypic traits (Doyle et al., 2017, Hedtke et al., 2017). The main difficulty associated with soft sweeps is the low degree of LD associated with any one locus that is under selection, thus requiring large sample size to detect the differences in allele frequency that results from the sweep. Interpreting the outcome of a soft sweep and designing an amplicon re-sequencing experiment, therefore requires an understanding of how much LD accompanies the soft sweep. These data highlight the importance of defining the degree of LD that is generated in the *O. volvulus* genome following soft and hard selection and that is the target of the next chapter in this thesis.

There are two good reasons for picking an amplicon in an amplicon resequencing study: (a) because it contains a causative SNP and/or (b) because it is in LD with the causative SNP. In this study, I selected high-F_{ST} regions for amplification that contains missense SNPs that could plausibly be either the causative SNPs or in LD with the causative SNP. Rather than following the hypothesis of finding the causative SNPs through amplicon resequencing, I took another approach. Other studies have been able to target causative SNPs for amplification, such as studying the mechanism of benzimidazole resistance in *H. contortus*. The β-tubulin locus is the target of benzimidazole anthelmintic drug in the parasite because of strong LD resulting from hard selection at the locus (Redman et al., 2015, Sallé et al., 2019). In contrast, in *O. volvulus* soft selective sweeps contribute to loss of drug sensitivity (Doyle et al., 2017, Hedtke et al., 2017, Gilleard and Beech, 2007). Therefore, in the next chapter, I estimated LD within a proposed QTL to explore how ivermectin response affects LD. This will enable me to select markers for GWAS that have a higher likelihood of being associated with the causative SNP.

The second and third criteria: Departures from neutrality (Tajima's D and Ka/Ks) - I surveyed signals that could support the hypothesis that selection on SOR is driving differentiation of the QTLs (as measured by F_{ST}). Tajima's D is a statistical tool to measure the difference between two estimators of the population mutation rate, θ_w (the product of the effective population size and the neutral mutation rate), and π, which measures the average number of pairwise differences between two DNA sequences (Tajima, 1989). Under neutrality, mean θ_w equals mean π (Tajima, 1989). The remarkable and important difference between θ_w and π is the effect of selection. Positive values of Tajima's D arise from an excess of intermediate frequency alleles and can result from population bottlenecks, population structure and/or balancing selection, while negative values of Tajima's D indicate an excess of low frequency alleles and can result from population expansions or positive selection (Tajima, 1989).

The Tajima's D values for each amplicon (as shown in Table 2.7) suggested that there are deviations from neutral evolution across some of the amplicons. For example, the Tajima's D value for OM4_SA9 amplicon, which has elevated F_{ST}, shows a weakly negative value (-0.163) which may indicate an excess of low frequency alleles resulting from population expansion or positive selection. The Ka/Ks value of 0 for the same amplicon does not support the Tajima's D and F_{ST} findings, rather, it indicates that non-synonymous mutations are absent. A value of 1 is suggestive of no selection under the appropriate framework. In contrast, the OM1b_SA16 amplicon, which did not show a high F_{ST}, did have a high Tajima's D (Tajima's D = 2.57; significant at $P < 0.05$), which indicates an excess allele frequency which can result from a population bottleneck and/or balancing selection. On the other hand, OM4_SA1 amplicon had a low Tajima's D (Tajima's D = -2.06; significant at $P < 0.05$) but has no high F_{ST} SNPs, which also indicates an excess of low frequency alleles resulting from population expansions or positive selection. In conclusion, there are two SNP loci on OM4 and OM1b chromosomes with significant values. Those loci show significant departures from neutrality, but the cause of that departure requires investigation.

Since all the selected amplicons are protein coding, the ratio of the number of nonsynonymous variants per nonsynonymous site and the number of synonymous variants per synonymous site (Ka/Ks) for those amplicons were evaluated to provide information about the evolutionary forces operating on the particular gene in them (Nei and Gojobori, 1986). For example, under neutrality Ka/Ks=1. For genes that are subject to functional constraint such that non-synonymous amino acid mutations are deleterious and purged from the population, Ka/Ks<1. The observation of Ka/Ks>1 provides evidence for positive selection (Nei and Gojobori, 1986, Suzuki and Gojobori, 1999). The value of Ka/Ks (1.848, not statistically significant) for amplicon OM1b_SA16 did not corroborate Tajima's D observation and indicates no selection. Also, the Ka/Ks value for the amplicon OM4_SA1 is 0.712, which also indicates no selection because it is not significant. Probably because there are so few mutations overall, Ka/Ks ratios are low. In summary, for Ka/Ks the conclusion is that there is no evidence for selection in any amplicon. This is not that surprising if there has been soft selection. Maybe longer evolutionary times (involving sufficient divergence for speciation) will provide more power to this test.

The fourth criterion (genetic association) - An additional test of the hypothesis that particular SNP loci are associated with SOR phenotype is to perform an association test using a linear regression analysis. Association analysis carried out between GR and SOR across the amplicons resulted in two SNP positions associated with SOR at $p < 1 \text{ X } 10^{-3}$ (OM4:4199425 and OM4:4433757) but not significant following Bonferroni correction. These loci were both located in the OM4_SA5 and OM4_SA6 amplicons, which do not have elevated-F_{ST} variants or significant deviation from 0 for Tajima's D and Ka/Ks<1. The Bonferroni correction (for multiple testing) is an extremely stringent test but it assumes the tests are independent. So, 1000 tests with p = 0.05 means 50 expected 'significant' (false-positive) tests even when there is no association. A Bonferroni corrected p-value of 0.05/1000 assumes all tests are independent. OM1a/b and OM4 loci are independent, and I show in the following chapter that LD is relatively low even between loci on the same chromosome (Chapter three, Results). The two loci (OM4:4199425 and OM4:4433757) in question are nearly 250kb apart, so it is very likely that the p-value of 10^{-3} is significant because those two loci are almost certainly segregating independently, that is, the genotype at one locus will not influence the genotype at the other locus, considering another scale (amplicon-wide) or using LD estimates to determine the number of independent tests (Goddard, 2009).

My conclusion based on the association study is that within OM4_SA5 amplicon, there is a gene that is plausibly causative with a synonymous SNP at position OM4:4199425 that could impact WBGene00246934 gene expression. WBGene00246934 is predicted to be among the G protein-coupled receptor (GPCR), rhodopsin-like superfamily. It is also an ortholog of *Brugia malayi* Protein Bm9770, isoform a (Choi et al., 2016). GPCRs are a large family of cell surface proteins that regulate many aspects of an organism's physiology and are important drug targets (Hauser et al., 2018). The Rhodopsin-like receptors are a family of proteins that comprise the largest group of the GPCRs. They represent a widespread protein family that includes hormone, neurotransmitter, and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins (Casey and Gilman, 1988). This is encouraging considering that the primary target of ivermectin is a ligand-gated channel at neuromuscular junctions (Cully et al., 1994). Similarly, within the OM4_SA6 amplicon, there is a gene that is plausibly causative, with a synonymous SNP at position OM4:4433757 that could impact WBGene00248067 gene with Haem peroxidase function. The molecular functions include peroxidase activity (response to oxidative stress) and haem binding (oxidation-reduction process), similar to an ortholog of *skpo-3* in *C. elegans* (Choi et al., 2016), which acts in host defence (Tiller, 2014). This gene may be involved in *O. volvulus* response to external triggers such as ivermectin. None of these genes are among the previously proposed genes of association with ivermectin response in either the GWAS or candidate gene studies in *O. volvulus* (Doyle et al., 2017).

2.5. **Conclusion.**

This is the first study to use targeted amplicon re-sequencing to verify SOR-QTLs observed in a GWAS of ivermectin response in *O. volvulus* and to survey other signals of selection for drug response in a worm population.

None of the missense (or nonsynonymous) SNPs that are identified in the original whole genome sequencing GWAS QTLs were in strong association with SOR following analysis of the amplicon re-sequencing data and therefore are unlikely to play a causative role in the SOR phenotype or to be predictive of ivermectin response in *O. volvulus*. However, the amplicon re-sequencing approach of short amplicons did identify four novel SNPs that were associated with SOR in *O. volvulus*. These four novel SNPs occur in genes that have not been previously identified in the literature as associated with ivermectin response in *O. volvulus*.

I recommend the need for further studies, particularly using larger amplicons, to validate more putative QTLs. Such further studies require increase in sample size for sufficient confidence in the SNPs or QTLs showing strong association with SOR. Alternatively, a cost-effective and less laborious methodology for GWAS might involve identifying a reduced marker set that could then be used in conjunction with genome imputation. In Chapter 3, I explored an alternative approach by analysing broad patterns of LD and haploblock structure across two autosomes in *O. volvulus*. Understanding LD in the worm is a necessary precondition for identifying the maximum distance between a genotyped SNP locus and an unknown causative variant that is needed to reliably detect a genetic association between the two. Similarly, understanding LD across the genome would help to understand mechanism of selection in the worm, for example, if QTLs associated with SOR correspond to regions of elevated LD, this would be an indicator of recent or current strong selection.

Chapter Three

Estimation of Linkage Disequilibrium in Onchocerca volvulus.

3.1. Introduction

The design and application of association studies have great potential to increase our understanding of the genetic architecture of complex traits such as drug resistance in nematodes (Doyle et al., 2017, Doyle and Cotton, 2019, Redman et al., 2015). Previous GWAS of ivermectin response in *O. volvulus* gave insights into the associations between genotype and sub-optimal response (SOR) in adult female worms from Ghana and Cameroon but with a limited number of low sequence coverage pool-seq worms, which can result into a stochastic variation in allele detection in the worms (Doyle et al., 2017). An additional GWAS was performed on genome sequences of individual worms, to identify regions with strong association with SOR phenotype in *O. volvulus* from Ghana but the number of worms analysed was limited because of the insufficient quality of the DNA extractions available (Hedtke et al., 2019). My study aims to develop methodologies that improve the power for GWAS in *O. volvulus*.

A particularly useful metric of LD is r^2, which is equivalent to the Pearson correlation coefficient between two markers (Hill and Robertson, 1968). r^2 is calculated by dividing *D* by the product of the four allele frequencies at the two loci being considered and ranges from 0 (no LD) to 1 (‘complete’ LD). Complete LD occurs if the markers have not been separated by recombination and therefore have the same allele frequency. In this case, observations at one marker provide complete information about the other marker, making the two redundant. r^2 has a simple inverse relationship with the sample size required to detect association between SNP loci, which makes it a commonly used metric to describe LD in genetic association studies (Weiss and Clark, 2002). For example, if one SNP (SNP1) is involved in disease susceptibility, but resistant and susceptible individuals are genotyped at a nearby site (SNP2), then sample size needs to be increased by a factor of $1/r^2$ to achieve the same power to detect association as the power if genotyping were done using SNP1 (Wall and Pritchard, 2003b). Generally, r^2-values above 0.33 are used to indicate strong LD sufficient for association studies, as the increase in sample size required to detect association should individuals not be genotyped at the causative loci is no more than threefold greater (reviewed in Ardie et al. (2002)). This concept was established based on the interpretation of r^2 in terms of its power to detect an association

(Kruglyak, 1999). Sample size is a limiting factor in association studies (Spencer et al., 2009), but large increases in sample size to compensate for weak LD between a marker and a causative QTL can be impractical. This is particularly true in studies that aim to use whole genome sequencing to identify QTLs associated with ivermectin response in *O. volvulus*, because of the limited amount of funding that is available for surgical removal of worms from people in the affected African countries, the labour and skill required to assess phenotype, and the reduced number of samples with genomic DNA of sufficient quality for sequencing (Doyle and Cotton, 2019, Hedtke et al., 2019). Values of $r^2 > 0.33$ limit the required increase in sample size to no more than threefold, and therefore, was considered to be the minimum useful LD values.

The spatial pattern of LD can also be measured by observing the haplotype block pattern in a genome. A definition of a haplotype block (called 'haploblock' in this study) proposed by Gabriel et al., is a set of consecutive loci (or SNPs) between which there is little or no evidence of historical recombination. That is, a haploblock is a set of closely linked loci on a chromosome with a strong LD between them so that they tend to be inherited together (Gabriel et al., 2002). By "strong LD", Gabriel et al. meant a one-sided upper 95% confidence bound on *D'* that is at least 0.98 and the lower bound above 0.7. Since *D'* values fluctuate upward with small sample size and/or presence of rare alleles, confidence bounds on *D'* were relied upon in defining haploblocks rather than point estimates (Wall and Pritchard, 2003b). In this study, this ensured that no pair of SNP loci has LD lower than an approximate *D'* of 0.7.

In addition to recombination between sites, LD (and by extension, haploblocks) is the result of the interaction of several possible biological and artifactual mechanisms, including population subdivision (which inflates LD), recurrent mutation, gene conversion, selection, or errors of genome assembly which disrupt haploblock patterns (Gabriel et al., 2002, Qanbari et al., 2010, Wall and Pritchard, 2003b). Locating haploblocks in the genome is of great practical importance for genetic association studies, to the extent that testing one SNP within each block for significant association with a trait might be sufficient to indicate association with every SNP in that block; that is, only one SNP in the haploblock may be required to identify an association of a particular genotype with a phenotype (Carlson et al., 2004). Similarly, the signatures of genomic regions under positive selection can be identified by studying the haploblock structure throughout the genome (Qanbari et al., 2010a). Identifying haploblocks makes it possible to predict the likely configurations of alleles at unobserved sites (Wall and Pritchard, 2003a).

There has been some (very limited) analysis of LD, specifically in nematodes. For example, a study by Cutter et al. revealed high nucleotide polymorphism and rapid decay of LD in wild populations of the free-living nematode *Caenorhabditis remanei* (Cutter et al., 2006). There have been very few studies in parasitic helminths. Doyle et al., reported a comprehensive analysis of genetic diversity and the construction of a F1 genetic map for the sheep gastrointestinal helminth *Haemonchus contortus* (Doyle et al., 2018), an LD study in *O. volvulus* based on short loci in P-glycoprotein gene was reported by Ardelli et al. (Ardelli et al., 2006), and Evans and Anderson (2020) used linkage mapping on recombinants derived from a cross between the laboratory strain and a wild strain to identify a single overlapping QTL on chromosome V that influences the responses of *C. elegans* to eight chemotherapeutic compounds. However, the extent of LD and haploblock structure has not been characterised before across an entire *O. volvulus* chromosome.

The objective of this study is to characterize patterns of LD decay and haploblock structure across the autosomes of the adult female *O. volvulus*. In this study, the criteria for measuring LD and haploblocks were applied to *O. volvulus* samples from Choi et al. (2016) and Hedtke et al. (2017). I applied the criteria to the study of LD in selected regions between sub-populations and across the length of two autosomes to identify if the same evolutionary force is driving both LD and genetic differentiation between good responders (GR) and suboptimal responders (SOR) to ivermectin. These two "sub-populations" (GR and SOR) are composed of worms that differ with respect to ivermectin response, but which are drawn from the same population of worms. I focused particularly on analysis of LD associated with regions of elevated genetic differentiation (F_{ST}) between GR and SOR worms from Ghana (Hedtke et al., 2017). F_{ST} (Wright's fixation index) is a measure of genetic differentiation between two populations (Holsinger and Weir, 2009) and will vary across the genome depending on whether the two sub-populations are genetically similar or genetically different in that region of the genome. The hypotheses tested were (1) that regions of elevated F_{ST} between GR and SOR will show higher LD than regions of lower F_{ST} (2) that regions of elevated FST between GR and SOR should also show elevated LD if there has been recent selection (especially a hard sweep), but that the LD may be weaker if the sweep is soft (3) that there will be regions of elevated LD that are not correlated with elevation of GR/SOR F_{ST} that are the result of selection that is not related to drug response. Then I generalised the LD analysis to encompass the whole of the autosomal genome without respect to ivermectin response and compared the LD and haploblock structure of the genome to the F_{ST} structure.

3.2. Materials and Methods

3.2.1. Study Data sets.

Processed reads (mapped data) of adult female *O. volvulus* (n = 96) from Ecuador, East Africa and West Africa (Table 3.1) were obtained from Choi et al. (2016) and S. Hedtke et al. (2017). A subset (n = 47) of the worms were divided up into phenotypes based on the continuing presence of microfilariae in the uteri of individual female worms after several rounds of ivermectin treatment in some study communities across Ghana (Osei-Atweneboana et al. 2011). In this study, the susceptible worms are referred to as good responders (GR: approximately 90 days after the person took ivermectin, there were no microfilariae in the uteri of the worm), while the resistant worms are referred to as sub-optimal responders (SOR: approximately 90 days after the person took ivermectin, there were microfilariae in the uteri of that female adult worm). Extraction of DNA from these worms, sequencing, read processing, mapping and variant calling were reported in previous studies (Armoo et al., 2017, Choi et al., 2016, Crawford et al., 2019; Hedtke et al. (in prep)). Variant calling and data processing was performed again with a similar unique pipeline. Knowing that among SNPs with higher MAF the estimate of LD tends to be stronger and that with higher MAF fewer SNPs are left for the estimation of LD, and this may introduce a bias (Yan et al. 2009). Variant sites supported by a minimum alternate count of 5, a minimum variant frequency of 0.25, a minimum depth of 20, and a minimum quality score of 30 were called from all worms using *Freebayes* v1.0.2 (Garrison and Marth, 2012).

3.2.2. Choice of Regions.

Regions of elevated F_{ST} that strongly differentiated GR worms from SOR worms were chosen based on an analysis carried out by S. Hedtke et al. (in prep) on a worm population from Ghana. Possible regions of interest were defined as regions with statistically significantly high F_{ST} using a cut-off value of 5 standard deviations from the mean.

To characterize LD systematically around those regions of elevated F_{ST} that differentiated GR and SOR sub-populations, three regions from chromosome OM1 of *O. volvulus* were examined (Figure 3.1). 1) ‘Region A’ (highlighted with blue arrow in Figure 3.1), comprises approximately 20 kb that surrounds a region on the genome with evidence for selection associated with drug susceptibility (‘high F_{ST} region’). 2) ‘Region B’ (highlighted with green arrow in Figure 3.1) is located approximately 50 kb away from region A. It is also approximately 20 kb in length and encompasses a region that has no association with drug resistance and that is presumably evolving neutrally (‘low F_{ST} region’). 3) ‘Region C’ (highlighted with brown arrow in Figure 3.1), is approximately 100 kb and includes regions A and B and the 50 kb gap in between.

Table 3.1. Study samples' location, country, and number of worm samples.

Location	**Country**	**Number of worms**
East Africa	Uganda	2
South America	Ecuador	10
West Africa	Benin	1
	Côte D'Ivoire	5
	Ghana	67
	Guinea	4
	Liberia	1
	Mali	5
	Sierra Leone	3
	Total	**98**

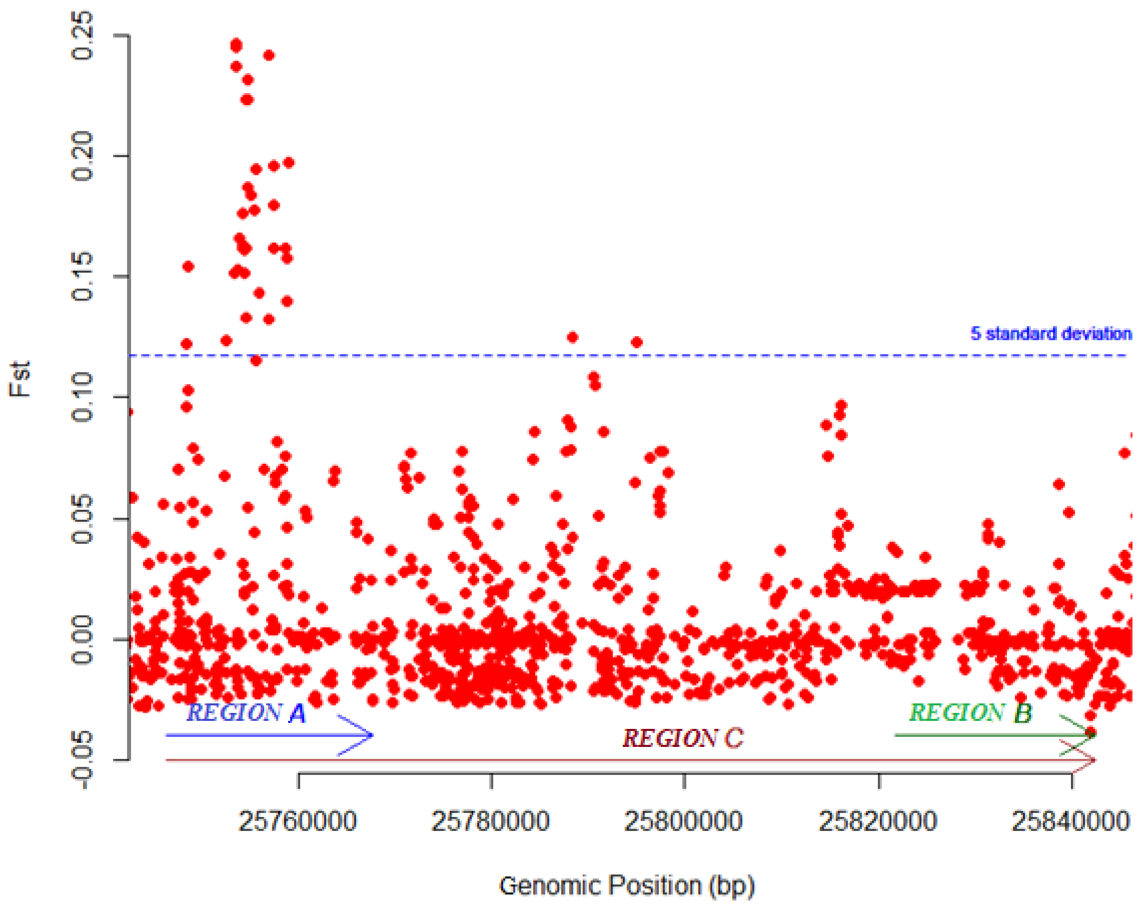


Figure 3.1. F_{ST} plot highlighting the regions of interest on chromosome OM1 of the *O. volvulus* genome.

F_{ST} plot showing the degree of genetic differentiation between two “sub-populations” (good responders (GR) and sub-optimal responders (SOR)), which are composed of worms that differ with respect to ivermectin response, but which are drawn from the same population of worms. The red dots are F_{ST} values at each SNP position. The blue dashed line shows the cut-off value for differentiation (+5 standard deviation). The blue arrow points to the location of the high F_{ST} region (Region A - showing strong differentiation between the GR and SOR), the green arrow points to the low F_{ST} locus (Region B - showing low differentiation between the GR and SOR), while the brown arrow points to Region C, which encompasses both high and low F_{ST} regions and the chromosomal region between them (Hedtke et al. 2017).

3.2.3. Measures of Linkage Disequilibrium.

The ideal scenario to measure the extent of LD within a population is to analyse non-related individuals. Previous studies also reported that both a multi-dimensional scale plot based on raw Hamming distance and a plot demonstrating the results of Discriminant Analysis of Principal Components (DAPC) (using 20 PCAs as per optimization in *pegas*) suggested that ivermectin-susceptible worms (GR) are not well differentiated from resistant worms (SOR) at a genomic level, and that resistant worms tend to represent a subset of the overall genetic diversity found in susceptible worms (Crawford et al., 2019, Doyle et al., 2017). Particularly, Doyle et al. used nuclear data and suggested that the Ghana transect is a single population and the phenotyped worms sequenced here were from the same population (also see Chapter 2 section 2.3 and Chapter 4 section 4.3 for tests of population structure on these data). Therefore, it is safe to assume that the samples used in this study are unrelated.

Linkage disequilibrium statistics based on adjacent SNP loci and pairwise SNP loci were estimated using *PLINK* version 1.09 (Purcell et al., 2007). All LD values for all the SNP loci within the chosen region of interest versus one specific SNP locus were estimated. The SNP locus has a criterion of high F_{ST} value (0.24) and high LD (r^2>0.8) for region A and C and low F_{ST} value (0.013) and high LD (r^2>0.8) for region B. The *PLINK –r2* command was used to estimate the correlations between each marker pair genome-wide within each sub-population as well. The average r^2 of adjacent SNP loci were estimated for each selected regions and chromosomes. Poisson model regression analysis was performed using the *nls* function in *Rstudio* (R Development Core Team, 2013) to fit the appropriate model for the rate of decay of LD for each region and within sub-populations. The allele frequencies were estimated using *Vcftools* v.0.1.3 (Danecek et al., 2011) across each sub-population.

To explore decay of LD across autosomes, the pairwise correlation of all SNP loci in windows of 100 kb intervals was measured across the entire lengths of the OM1 and OM4 chromosomes using *PLINK* version 1.09 (Purcell et al., 2007). Pairwise LD between SNP loci were grouped by their pairwise physical distance into intervals of 1 kb. Average r^2 for SNP pairs in each interval was estimated and compared with the respective average distance between the SNP pairs.

3.2.4. Haplotype block estimation across the genome.

PLINK version 1.09 (Purcell et al., 2007) was used to define the haploblocks present in the regions of specific interest and across chromosomes OM1 and OM4. The method followed for block definition has been previously described by Gabriel et al. (2002) and described in the introduction to this chapter. Block size was limited using the command '--blocks-max-kb'. Maximum block sizes of 20 kb were considered for regions A and B, while maximum block sizes of 100 kb were considered for region C.

All figures and tables were created in *R Studio* using the *ggplot2* package (Wickham, 2016) and Microsoft Excel (Microsoft Corporation, Redman, Washington, USA).

3.3. Results

3.3.1. Summary statistics of the variants.

I explored LD across three regions on the OM1 chromosome of *O. volvulus* using the whole data sets regardless of sub-population in order to maximize sample size. The distribution of SNPs per region is described in Table 3.2. In total, 676 variants, 375 variants, and 2699 variants were screened across 98 adult female worms in regions A (high F_{ST} region), B (low F_{ST} region), and C, respectively. The average estimated physical distance between adjacent SNP loci was approximately 43 bp for region A, 76 bp for region B, and 51 bp for region C. The longest interval between adjacent SNP loci was 424 bp on region A, 643 bp on region B, and 1,094 bp on region C (Table 3.2).

Considering the impact of subpopulation structure, LD between two variant sites was measured using r^2 for 47 unrelated adult female phenotyped *O. volvulus* worms from Ghana to identify if similar evolutionary forces are operating at these regions among the GR and SOR worms and to assess if the genetic differentiation observed (from previous genome wide Fst analysis) was real or an artefact of other underlying selection. The summary statistics of the variants used in downstream analysis is described in Table 3.2. The table shows that 346 variants were called across 25 unrelated GR worms (region A (high F_{ST} region) variants = 191, region B (low F_{ST} region) variants = 155) and 295 variants across 22 unrelated SOR female worms (region A (high F_{ST} region) variants = 210, region B (low F_{ST} region) variants = 85) (Table 3.2). The average estimated physical distance between adjacent SNP loci was similar for region A (high F_{ST} region) and B (low F_{ST} region) in GR sub-population (approximately 109 bp versus 111 bp). The average estimated physical distance between adjacent SNP loci for region B (low F_{ST} region) in SOR sub-population is almost three times that of region A (high F_{ST} region) (228 bp versus 98 bp). The longest interval between adjacent SNP loci is 2192 bp on region A and 773 bp on region B (low F_{ST} region) among the GR sub-population, and 2078 bp on region A (high F_{ST} region) and 2778 bp on region B (low F_{ST} region) among the SOR sub-population.

Average LD between adjacent SNP loci across OM1 and OM4 chromosome of the *O. volvulus* genome was also estimated and is shown in Table 3.2. The number of SNP loci was in proportion to the chromosome length (422,125 in OM1 and 233,375 in OM4), providing an SNP density of 13.95 and 14.95 per kb (one SNP per 74 and 68 bp average) across OM1 and OM4, respectively (Table 3.2). The average estimated physical distance between adjacent

SNP loci was approximately 74 bp and 62 bp for OM1 and OM4, respectively. The longest interval between adjacent SNP loci was 100.166 Kb on OM1 and 200.967 kb on OM4.

Table 3.2. A summary of variants used in downstream analysis.

Phenotype		No of individuals	Length (bp)	No of SNPs	Average Distance between adjacent SNP loci (bp)	r^2 for all pairs of adjacent SNPs	Average Distance between all Pairwise SNPs (kb)	r^2 for all pairwise combinations	MAF
GR	**region**					**Mean (SD)**		**Mean (SD)**	**Mean (SD)**
	region A	25	21,368	191	108	0.23(0.35)	6.64	0.10 (0.19)	0.12 (0.14)
	region B	25	20,623	155	111	0.40(0.43)	7.05	0.23 (0.33)	0.17 (0.21)
SOR	**region**								
	region A	22	21,368	210	98	0.21(0.34)	6.74	0.13 (0.19)	0.17 (0.15)
	region B	22	20,623	85	228	0.32(0.39)	6.33	0.23 (0.28)	0.27 (0.20)
Combined dataset									
	region A	98	21,368	676	43	0.23(0.35)	7.17	0.11 (0.16)	0.14 (0.12)
	region B	98	20,623	375	76	0.45(0.44)	6.85	0.17 (0.26)	0.15 (0.17)
	region C	98	96,004	2699	51	0.31(0.40)	38.59	0.10 (0.16)	0.13 (0.14)
Chromosome									
	OM1	98	31,161,767	422,125	74	0.25 (0.25)	487.34	0.08 (0.03)	0.18
	OM4	98	16,048,563	233,375	62	0.21(0.31)	497.91	0.08(0.02)	0.18

NB: region A = High F_{ST} region (Length = 20 kb); region B = Low F_{ST} region (Length = 20 kb); region C = Combined both regions A and B (Length = approximately 100 kb); GR = "susceptible worms"; SOR = "resistant worms"; No of SNPs = number of SNP loci called per region; Average physical distance between adjacent SNP loci (bp) = average separation between adjacent SNP loci; Average Distance between all Pairwise SNPs = the average distance between SNPs across all pairwise combinations; r^2 for all pairwise combinations = Pairwise Linkage disequilibrium measures; r^2 for all pairs of adjacent SNPs = Linkage disequilibrium between SNPs that are not more than 2 SNPs apart; Mean (SD) = Mean and standard deviation; MAF = Minor Allele Frequency

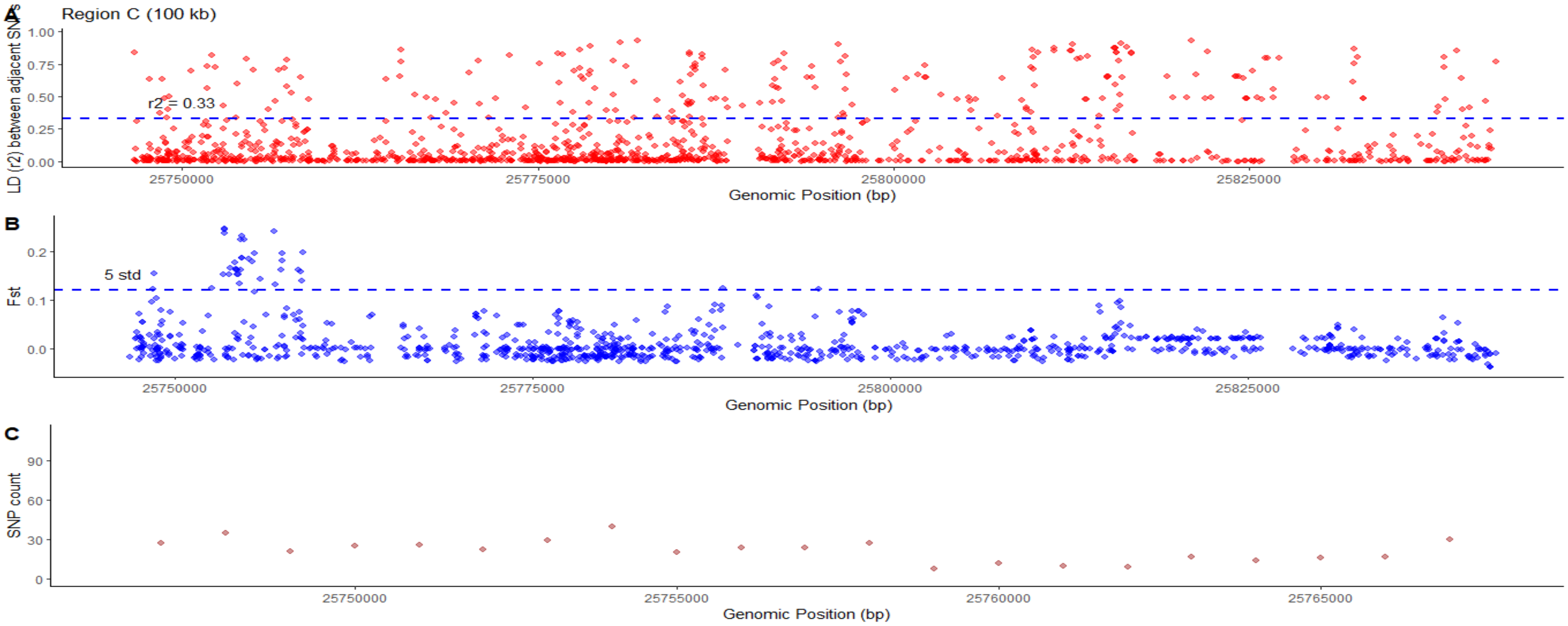


Figure 3.2. Linkage disequilibrium, F_{ST} and SNP density between adjacent SNP pairs of a 100 Kb region (region c) on OM1 chromosome.

The X-axis is the genomic positions across a randomly selected 100 Kb region on OM1 chromosome, while the Y-axis is A) the distribution of LD between adjacent SNP pairs, B) the degree of differentiation (F_{ST}) between susceptible (GR) and resistant (SOR) worms, and C) the density of SNP across the region binned in a 1 kb interval. The blue dotted horizontal line in plot A is the minimum r^2 values regarded as useful LD for association studies, while the blue solid line on plot B shows the cut-off point of the F_{ST} differentiation (5 standard deviation).

NB: Only values of r^2 between 0 and 0.99 were used because values with $r^2 = 1$ shows that no recombination is occurring at those sites

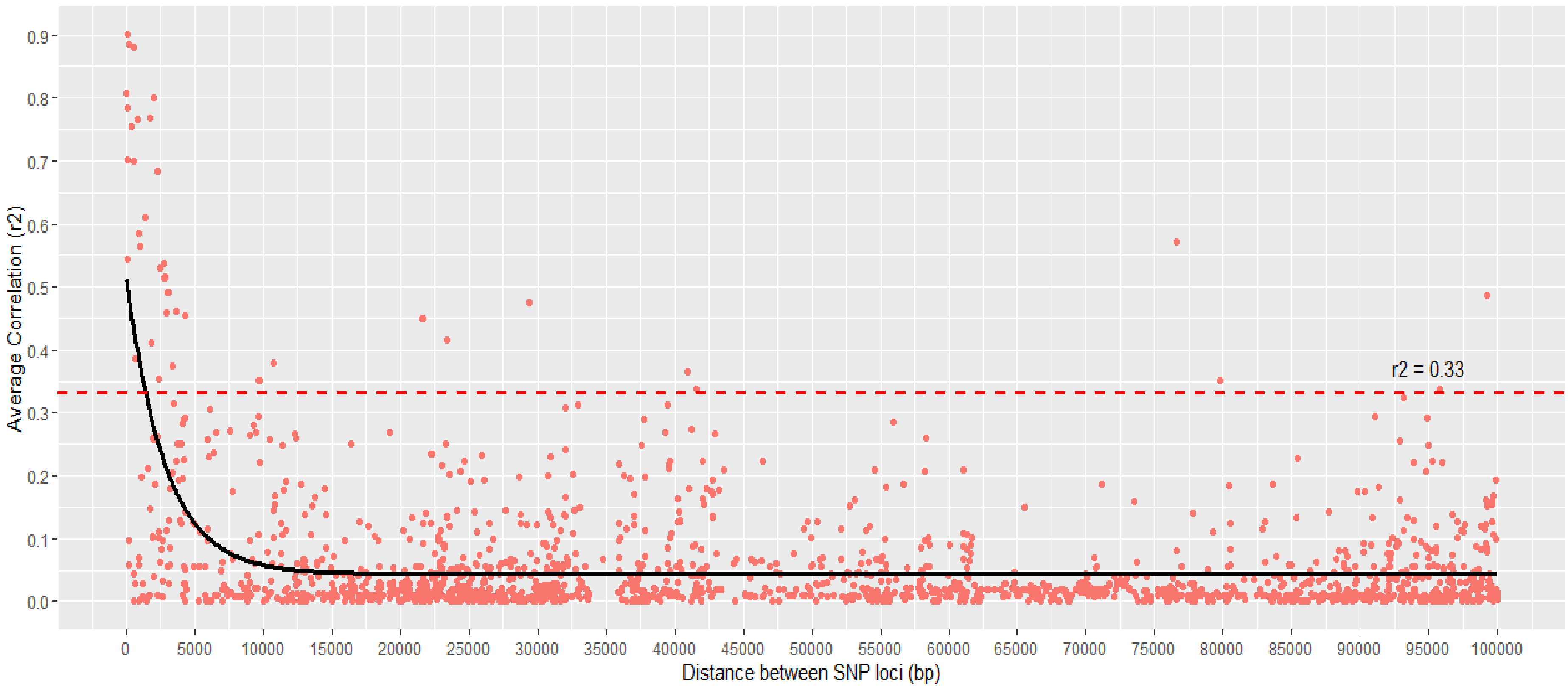


Figure 3.3. Graph of the predicted decline in linkage disequilibrium (r^2) with distance within an approximate 100kb region (region c) from a QTL in *O. volvulus* genome.

Plot showing LD values for every SNP locus within 100 kb of a SNP of interest within the QTL (region C). X-axis: the distance between SNP loci. Y-axis: the measure of linkage disequilibrium (r^2). The dashed red line is the useful LD threshold at r^2 = 0.33. Sample size of 98 adult female *O. volvulus* worms were used. The graph overlays the lines of expected that is the 'modelled' values (black line) onto the actual points (dots). The figure shows an asymptotic decline in LD with increasing separation between markers.

3.3.2. Linkage Disequilibrium

3.3.2.1. Estimating linkage disequilibrium over 100 kb of chromosome OM1 corresponding to the QTL region of interest.

LD was calculated over a 100 kb region on OM1 chromosome of the *O. volvulus* using r^2. The whole data set (n = 98) regardless of sub-population was used to maximize sample size. Figure 3.2A plots all values of r^2 for each adjacent locus pair across a 100kb region, while Figure 3.3 plots the values of r^2 of a locus of interest (within a QTL) with other SNP loci within a 100 kb region to its right (that is, all LD values for every SNP within 100 kb of a SNP of interest within the QTL). The "SNP-of-interest" was chosen using criteria of high F_{ST} value (0.24) and high LD (r^2>0.8)). Non-linear regression analysis was performed using the *nls* function to fit the appropriate exponential model for the rate of LD decay in region C as seen in figure 3.3. The maximum slope for the model is -0.00017 (negative sign indicating decay) where the distance between SNPs is 1 and the LD is 0.52 (y0, intercept parameter on the output; Std error = 0.03786). Then the slope keeps decreasing till it asymptotes at around ~10,000 bp towards 0.048 (yf) at a rate alpha (Std error = 0.01117). As expected, the exponential curve model on Figure 3.3 shows that there is an asymptotic decline in LD with increasing separation between SNP loci. Figure 3.3 shows that LD started at approximately 0.52, decayed rapidly to 0.33 (the threshold for useful LD estimate) at a distance of approximately 1405 bp, then extended gradually to an asymptote level of 0.05 at approximate distance of 9-10 kb (which shows the point at which linkage equilibrium is reached, that is, where there is no correlation between SNP loci because they are too far apart). All pairwise LD combinations for pairs of SNP loci within an approximately 100 kb region (region C) was also estimated and the average LD per bin distance of 1 kb in that region ranged from 0.07 to 0.18 (Table 3.3).

The two 20 kb regions at either end of region C are denoted as regions A and B, which correspond to regions of high and low F_{ST}, respectively. Figure 3.4 shows the decline of LD over those regions while figure 3.5A and B plots all values of r^2 for each adjacent locus pair across at those regions. Non-linear regression analysis was also performed using the *nls* function to fit the appropriate exponential model for the rate of LD decay in regions A and B as seen in figure 3.4. As expected, the exponential curve model on Figure 3.4 shows that there is an asymptotic decline in LD with increasing separation between SNP loci. Region B contains a small number of SNP combinations with elevated LD, so that the LD decline was predicted to start at 0.33 and declines rapidly to 0.2 at approximately 2kb distance then continues to the asymptote levels (r^2 = 0.05) by approximately 8-9 kb distance. In region A, the decline in LD is like that of region C but starts from a higher level (presumably reflecting the presence of a candidate QTL or as a result from lower recombination in that region because lower recombination would favour reduced gene flow and hence higher F_{ST}, even in the absence of a QTL.), so that LD starts at an approximate value of 0.52, then declines rapidly to 0.33 at an approximate distance of 1405 bp and reaches an asymptote level of approximately 0.05 at 9-10kb distance.

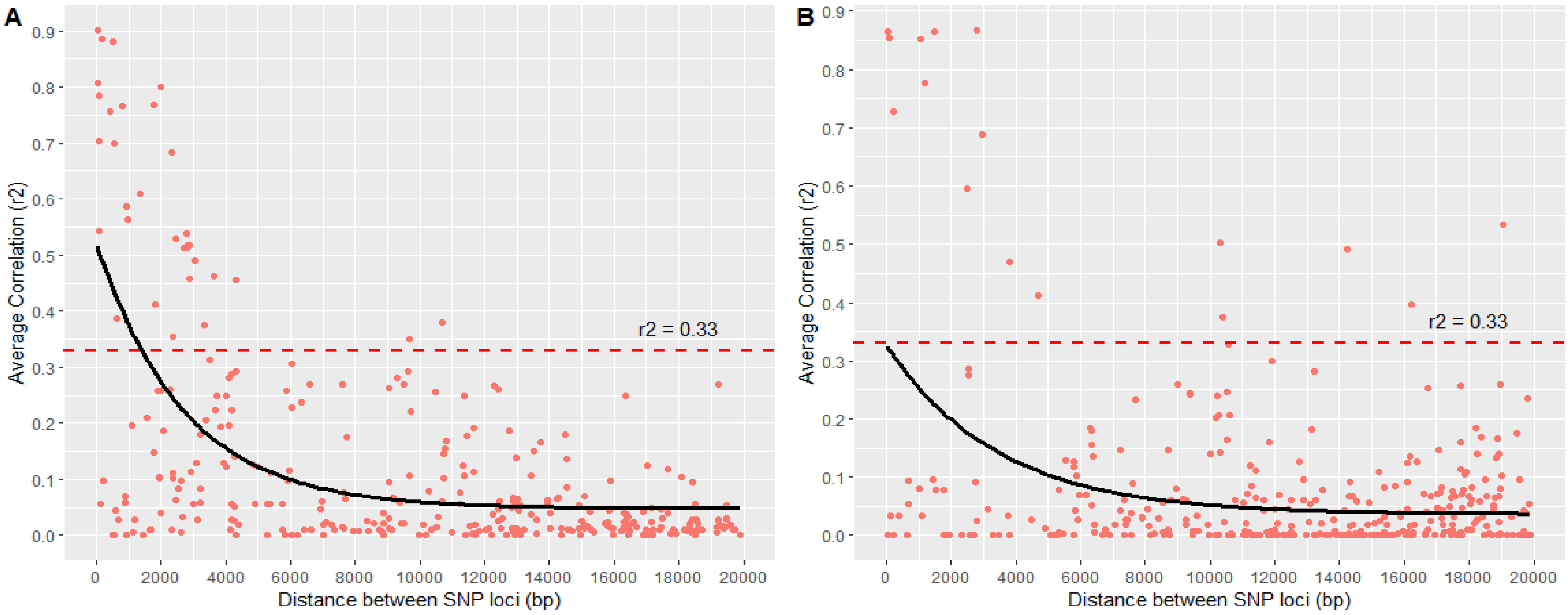


Figure 3.4. Graph of the predicted decline in linkage disequilibrium (r^2) with distance across all worms in two 20 kb regions on the *O. volvulus* genome.

Plot showing LD values for every SNP within 20 kb of a SNP of interest within (A) the QTL (region A) and (B) outside a QTL (region B). X-axis: the distance between SNP loci. Y-axis: the measure of linkage disequilibrium (r^2). The dashed red line is the useful LD threshold at $r^2 = 0.33$. Sample size of 98 adult female *O. volvulus* worms were used for all regions. The graph overlays the lines of modelled values (black line) onto the actual points (dots).

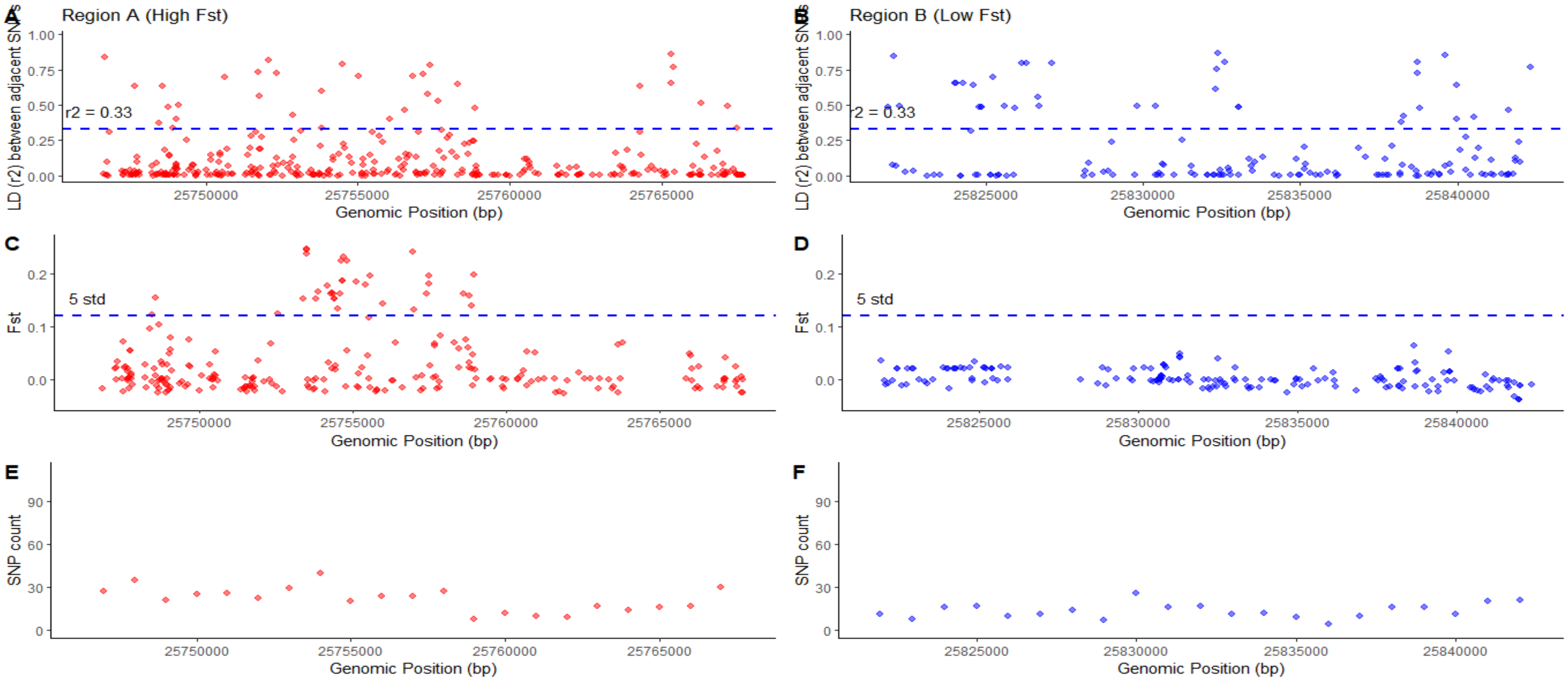


Figure 3.5. Linkage disequilibrium, F_{ST} and SNP density between adjacent SNP pairs in two selected 20 kb regions on OM1 chromosome across all worms.

LD for pairs of adjacent SNP loci within two approximately 20 kb regions was calculated as the squared correlation coefficient (r^2). The X-axis is the genomic positions across a randomly selected 20 kb regions on the OM1 chromosome for (A) Region A (high F_{ST} region) and (B) Region B (low F_{ST} region), while the Y-axis is the A) the distribution of LD between adjacent SNP loci, B) the degree of differentiation (F_{ST}) between susceptible (GR) and resistant (SOR) worms, and C) the density of SNP across those regions binned in a 1 kb interval. The blue dotted horizontal line in plot A is the minimum r^2 values regarded as useful LD for association studies, while the blue solid line on plot B shows the cut-off value for significant F_{ST} differentiation (5 standard deviations).

NB: Only values of r^2 between 0 and 0.99 were used because values with r^2 = 1 shows that no recombination is occurring at those sites

For association studies, LD estimates where $r^2 > 0.33$ are considered likely to be informative. There are 531 adjacent SNP loci in the 100kb interval defined as region C that meet this threshold, representing 33.50% of loci across the region (Figure 3.2). The average physical separation between adjacent SNP loci where LD meets this threshold is 42 bp (the largest interval for which r^2> 0.33 is 1,094 bp) (Figure 3.5A). Useful LD estimates are higher in region B (the low F_{ST} region), where 149 of adjacent SNP loci have LD (r^2) > 0.33 (65.64%) (Figure 3.5B). For region A, 92 pairs of SNP loci (22.38%) meet this criterion. The average physical separation between adjacent SNP loci where LD meets this threshold is 63 bp (the largest interval for which r^2> 0.33 is 334 bp) in region B compared to 28 bp in region A (the largest interval for which r^2> 0.33 is 217 bp).

Pairwise LD for pairs of SNP loci within region A and B were estimated and the average LD per bin distance of 1 kb in those regions ranged from 0.08 to 0.17 in region A (high F_{ST} region) and 0.09 to 0.26 in region B (low F_{ST} region) (Table 3.4). There is higher average all pairwise combinations LD in region B in comparison with A and C is being driven by a relatively small number of closely spaced SNP pairs with elevated LD (hence the greater range of pairwise LD values; Table 3.4). Surprisingly, the apparent maximum pairwise LD value in region B was not detected in the overlapping region C at the equivalent position.

Table 3.3. Mean linkage disequilibrium region C SNPs over different map distance.

Bin distance (bp)	*R2*
	Mean (SD)
< 1kb	0.181(0.263)
1 – 2 kb	0.145(0.222)
2 - 3 kb	0.131(0.207)
3 – 4 kb	0.126(0.199)
4 – 5 kb	0.12(0.192)
5 – 6 kb	0.117(0.191)
6 – 7kb	0.113(0.183)
7 – 8 kb	0.106(0.18)
8 – 9 kb	0.103(0.176)
9 – 10 kb	0.102(0.176)
10 – 11 kb	0.1(0.173)
11 – 12 kb	0.094(0.164)
12 – 13 kb	0.096(0.162)
13 – 14 kb	0.096(0.163)
14 – 15 kb	0.094(0.162)
15 – 16 kb	0.093(0.161)
16 – 17 kb	0.095(0.161)
17 – 18 kb	0.091(0.155)
18 – 19 kb	0.094(0.157)
19 – 20 kb	0.096(0.157)
20 – 21 kb	0.09(0.155)
21 – 22 kb	0.089(0.155)
22 – 23 kb	0.087(0.152)
23 – 24 kb	0.09(0.154)
24 – 25 kb	0.091(0.157)
25 – 26 kb	0.089(0.156)
26 – 27 kb	0.088(0.157)
27 – 28 kb	0.083(0.15)
28 – 29 kb	0.082(0.146)
29 – 30 kb	0.084(0.147)
30 – 31 kb	0.083(0.148)
31 – 32 kb	0.082(0.148)
32 – 33 kb	0.081(0.145)
33 – 34 kb	0.083(0.148)
34 – 35 kb	0.082(0.146)
35 – 36 kb	0.087(0.15)
36 – 37 kb	0.086(0.152)
37 – 38 kb	0.083(0.149)
38 – 39 kb	0.087(0.154)
39 – 40 kb	0.09(0.157)
40 – 41 kb	0.088(0.157)
41 – 42 kb	0.084(0.15)
42 – 43 kb	0.082(0.141)
43 – 44 kb	0.085(0.15)
44 – 45 kb	0.085(0.148)
45 – 46 kb	0.085(0.149)
46 – 47 kb	0.077(0.144)
47 – 48 kb	0.077(0.145)
48 – 49 kb	0.078(0.144)
49 – 50 kb	0.081(0.151)
50 – 51 kb	0.08(0.15)
51 – 52 kb	0.077(0.151)
52 – 53 kb	0.075(0.147)
53 – 54 kb	0.075(0.147)
54 – 55 kb	0.077(0.147)
55 – 56 kb	0.075(0.144)
56 – 57 kb	0.076(0.14)
57 – 58 kb	0.085(0.15)
58 – 59 kb	0.079(0.141)
59 – 60 kb	0.078(0.139)
60 – 61 kb	0.079(0.136)
61 – 62 kb	0.077(0.133)
62 – 63 kb	0.081(0.14)
63 - 64 kb	0.084(0.148)
64 – 65 kb	0.085(0.154)
65 – 66 kb	0.083(0.151)
66 – 67 kb	0.079(0.15)
67 – 68 kb	0.076(0.142)
68 – 69 kb	0.077(0.142)
69 – 70 kb	0.077(0.142)
70 – 71 kb	0.072(0.143)
71 – 72 kb	0.071(0.138)
72 – 73 kb	0.072(0.14)
73 – 74 kb	0.072(0.136)
74 – 75 kb	0.072(0.133)
75 – 76 kb	0.067(0.126)
76 – 77 kb	0.076(0.141)
77 – 78 kb	0.071(0.137)
78 – 79 kb	0.077(0.14)
79 – 80 kb	0.076(0.131)
80 – 81 kb	0.071(0.13)
81 – 82 kb	0.076(0.136)
82 – 83 kb	0.072(0.131)

83 – 84 kb	0.07(0.127)
84 – 85 kb	0.074(0.134)
85 – 86 kb	0.075(0.128)
86 – 87 kb	0.074(0.131)
87 – 88 kb	0.081(0.134)
89 – 90 kb	0.08(0.135)
90 – 91 kb	0.078(0.135)
91 – 92 kb	0.079(0.139)
92 – 93 kb	0.073(0.123)
92 – 93 kb	0.069(0.118)
93 – 94 kb	0.082(0.131)
94 – 95 kb	0.067(0.107)
95 – 96 kb	0.107(0.137)

Table 3.4. Mean linkage disequilibrium among high and low F_{ST} SNPs over different map distances.

Distance	Region A (High F_{ST})	Region B (Low F_{ST})
	Mean (SD)	Mean (SD)
< 1 kb	0.181 (0.250)	0.293 (0.357)
1 - 2 kb	0.135 (0.188)	0.210 (0.289)
2 - 3 kb	0.122 (0.175)	0.210 (0.278)
3 - 4 kb	0.118 (0.165)	0.199 (0.257)
4 - 5 kb	0.107 (0.147)	0.181 (0.245)
5 - 6 kb	0.099 (0.141)	0.197 (0.301)
6 - 7 kb	0.093 (0.135)	0.171 (0.268)
7 - 8 kb	0.095 (0.141)	0.151 (0.245)
8 - 9 kb	0.086 (0.127)	0.157 (0.255)
9 - 10 kb	0.094 (0.135)	0.129 (0.196)
10 - 11 kb	0.099 (0.143)	0.116 (0.178)
11 - 12 kb	0.089 (0.127)	0.107 (0.168)
12 - 13 kb	0.092 (0.126)	0.097 (0.156)
13 - 14 kb	0.099 (0.142)	0.095 (0.140)
14 - 15 kb	0.095 (0.139)	0.0942 (0.157)
15 - 16 kb	0.095 (0.144)	0.074 (0.130)
16 - 17 kb	0.087 (0.124)	0.074 (0.130)
17 - 18 kb	0.068 (0.095)	0.081 (0.150)
18 - 19 kb	0.074 (0.115)	0.095 (0.152)
19 - 20 kb	0.069 (0.102)	0.106 (0.157)

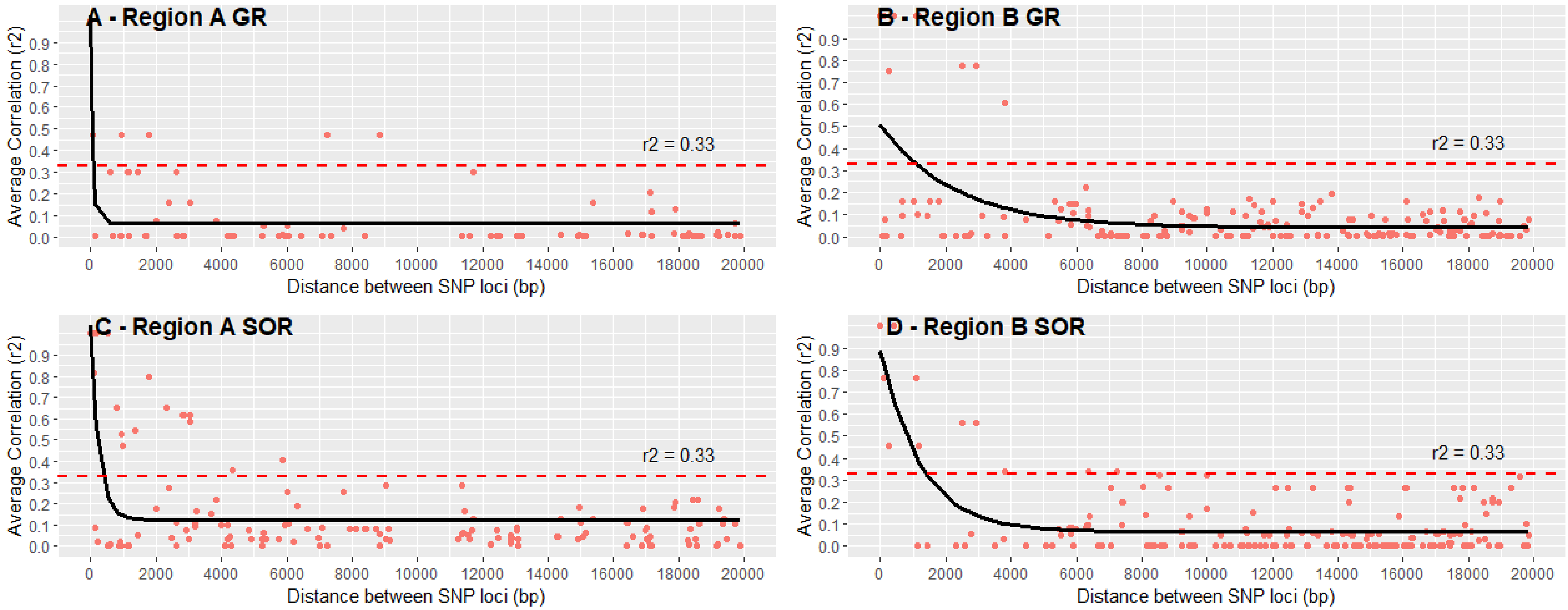


Figure 3.6. Graph of the predicted decline in linkage disequilibrium (r^2) with distance across two 20 kb regions among sub-populations of *O. volvulus*.

Plot showing LD values for every SNP within 20 kb of a SNP of interest in (A) within the QTL (region A) in susceptible worms (GR; n=25); (B) outside a QTL (region B) in susceptible worms (GR; n=25); (C) within the QTL (region A) in resistant worms (SOR; n=22), and (D) outside a QTL (region B) in resistant worms (SOR; n=22) of *O. volvulus* from Ghana. The graph overlays the lines of expected (modelled) values (black lines) onto the actual points (dots). The dashed red line is the useful LD threshold at $r^2 = 0.33$. The graph indicates that the expected rate of r^2 decline is faster in region A (high F_{ST} region) compared to region B (low F_{ST} region) within both sub-populations. The figure indicates that the decay in LD is dependent on the region investigated rather than the worm population.

3.3.2.2. Linkage disequilibrium estimate by sub-populations over the selected 20 kb regions.

Under the hypothesis that selection will have effect on LD, I compared the decline of LD between SNP loci using r^2 in the SOR (that is, worms selected for by drug) versus the GR (that is, worms not selected for by drug) around the candidate QTL in region A (high F_{ST}) and outside a QTL in region B (low F_{ST}). Figure 3.6 shows the graph of the predicted decline in LD (r^2) with distance among the *O. volvulus* sub-populations from Ghana. The figure reveals a difference in LD decay between region A (high F_{ST}) and region B (low F_{ST}) within the two sub-populations. Non-linear regression analysis was also performed using the *nls* function to fit the appropriate exponential model for the rate of LD decay in regions A and B within sub-populations as seen in figure 3.6. As expected, the exponential curve model on Figure 3.6 shows that there is an asymptotic decline in LD with increasing separation between SNP loci. Region A in GR and SOR sub-populations (Figure 3.6A&C) show similar starting point of predicted LD but reaches equilibrium at different rates and distances. LD decline was predicted to start at 1.0 and declines rapidly to 0.33 (the threshold for useful LD estimate) at a distance of approximately 200 bp in GR and a distance of approximately 500 bp the SOR sub-population, then continues to the asymptote level (r^2 = 0.05) at approximately 1 kb distance in the GR and to an asymptote level (r^2 = 0.1) at approximately 1.5 kb distance in the SOR. LD was predicted to start at a higher point (r^2 = 0.88) in region B in the SOR sub-population (Figures 3.6 D) compared to region B in the GR sub-population (r^2 = 0.52) (Figure 3.6 B), then declines to 0.33 (the threshold for useful LD estimate) in the GR sub-population at approximately 1kb distance compared to SOR (approximate distance of 1.5kb) and reaches an asymptote level (r^2 = 0.05) faster at approximately 5kb in SOR sub-population compared to approximately 6 kb in the GR sub-population.

Overall, a faster decline in LD to the threshold for useful LD was observed in the GR sub-population compared to the SOR in regions A (high F_{ST}) but LD reaches equilibrium faster in the GR sub-population compared to SOR in region A (high F_{ST}). On the other hand, in region B (low F_{ST}), LD started higher and reaches equilibrium faster in the SOR sub-population compared to GR.

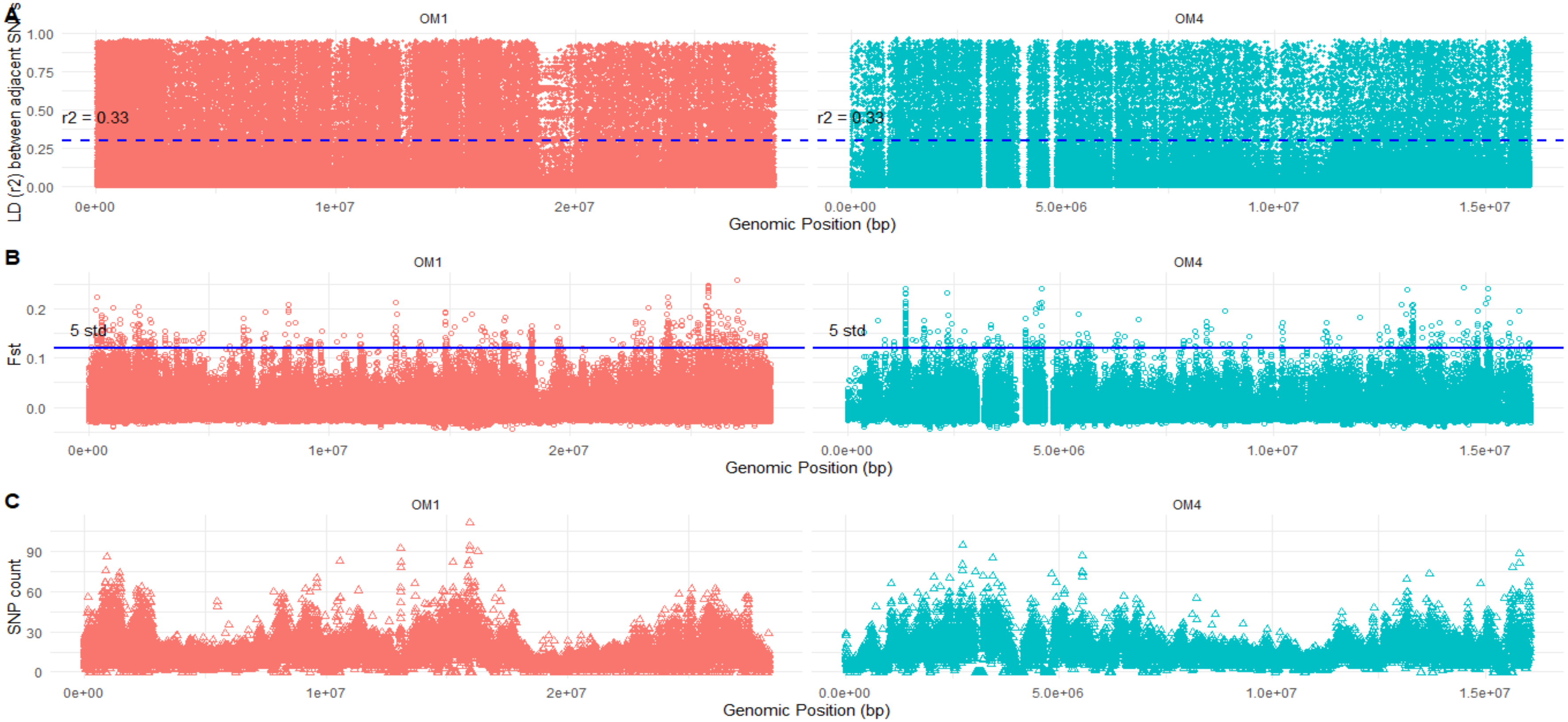


Figure 3.7. Linkage disequilibrium, F_{ST} and SNP density between adjacent SNP pairs of two *O. volvulus* autosomes.

Linkage disequilibrium for pairs of adjacent SNP loci on two *O. volvulus* autosomes were calculated as the squared correlation coefficient (r^2). The X-axis is the genomic positions across the chromosome for (A) OM1 and (B) OM4, while the Y-axis is the A) the distribution of LD between adjacent SNP pairs, B) the degree of differentiation (F_{ST}) between susceptible (GR) and resistant (SOR) worms, and C) the density of SNP across those regions binned in a 1 kb interval. The blue dotted horizontal line in plot A is the minimum r^2 values regarded as useful LD for association studies, while the blue solid line on plot B shows the cut-off point of the F_{ST} differentiation (5 standard deviation).

NB: Only values of r^2 between 0 and 0.99 were used because values with $r^2 = 1$ means no recombination is occurring at those sites

3.3.2.3. Linkage Disequilibrium estimates across two *O. volvulus* autosomal chromosomes.

Linkage disequilibrium decay was estimated across the entire length of two autosomes (OM1 and OM4) of *O. volvulus* to identify the pattern of LD decay across these chromosomes. At the time of this analysis, chromosome OM1 had not been fully assembled and was in two contigs OM1a and OM1b. Subsequent research (Cotton et al., 2016) used long read sequencing to determine that these two contigs are separated by minimum of 50 kb of repetitive sequence. My assumption is that LD between variants located on these two contigs is not significant to this analysis despite their being on the same chromosome, because recombination is likely to have occurred frequently between them given the extensive size of this repeat region.

Generally, LD was variable across the chromosome as indicated in Figure 3.7, where LD between adjacent SNP loci was compared with F_{ST} and SNP density across the chromosome length. The average and standard deviations for r^2 between adjacent SNP loci are approximately 0.25 (0.25) and 0.21 (0.31) for chromosome OM1 and OM4, respectively. LD estimates may be more frequent on chromosome OM1 (the number of adjacent SNP loci having LD (r^2) > 0.33 is approximately 89,513 (26.48%) SNP loci) compared to chromosome OM4 (approximately 47,470 (22.29%) adjacent SNP loci). The average physical separation between adjacent SNP loci where LD met this threshold is 56 bp in OM1 (with maximum physical separation of 4.994 kb) and it is 57 bp in OM4 (with maximum physical distance of 200.967 kb).

Figure 3.8 A and B shows what average LD between all pairwise combinations look like in a 1 Mb window across the autosomes of *O. volvulus*. On the average, two loci that are closer together should have high LD compared to those that are farther apart. The starting points of average LD on these plots were quite low because they were binned in 1kb intervals. The figure shows that LD decline with distance. In some windows on the OM1b chromosome, LD decay started at average r^2 greater than 0.300 (for example, windows between 13-14 Mb and 19-20 Mb on OM1b chromosome started at r^2 = 0.331 and 0.303, correspondingly), while some started at a lower value than that (for example, in windows between 26-27 Mb of the chromosome, average r^2 started at 0.150 (Figure 3.8A). On all windows in chromosome OM1a and OM4, the starting point of LD was at r^2 value of 0.14 (Figure 3.8B). In all, each 1Mb interval behaved in much the same way, with some variation in (a) the max LD value

observed and (b) the steepness of the decay. All eventually asymptote to approximately the same value (r^2 = 0.08). However, some of the LD curves behaved differently (increase at longer range). Also, it is expected that LD breaks down quickly at the edge of the chromosome as observed in windows 26-27 Mb and 27-28Mb, respectively.

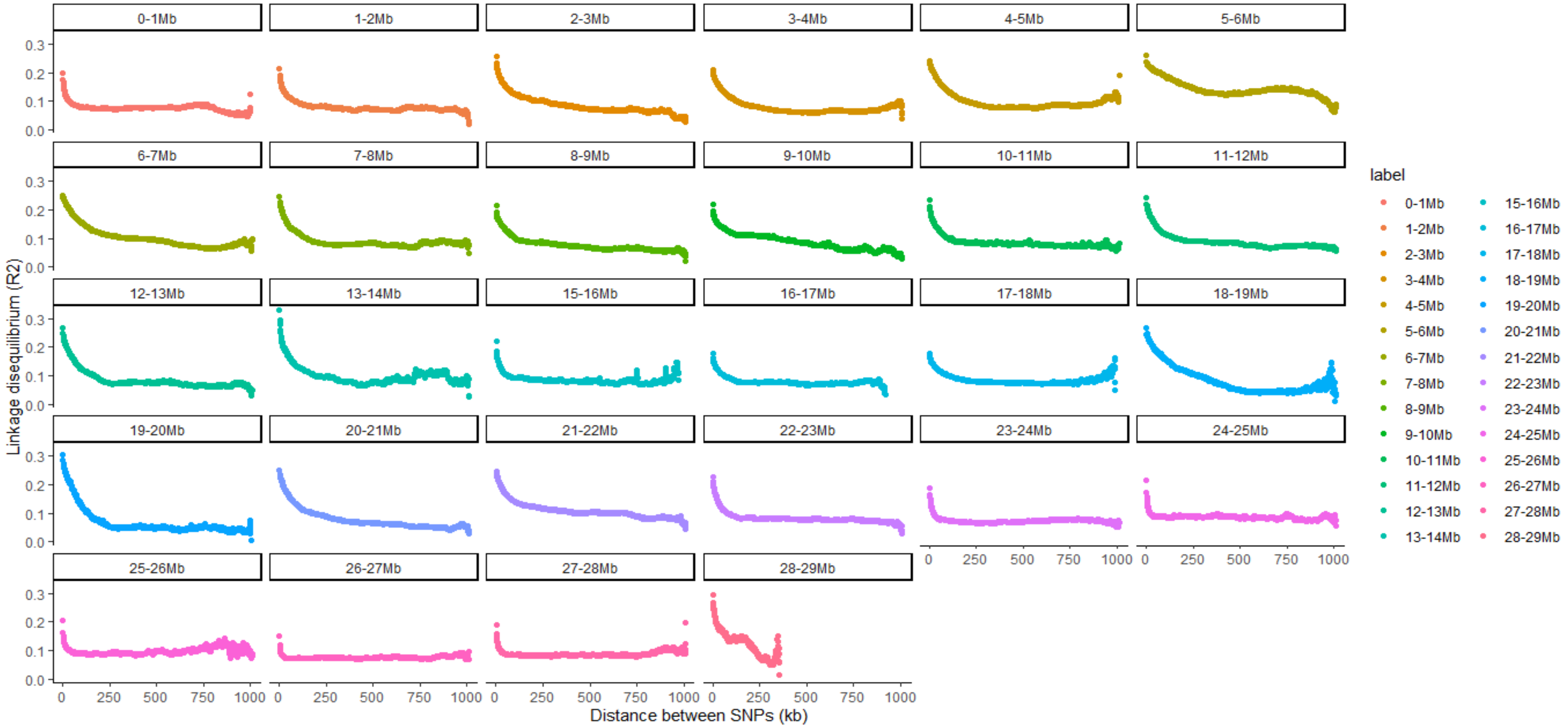


Figure 3.8 a. Linkage Disequilibrium (LD) decay plots for chromosome OM1 divided into 1 Mb window

Average LD for all pairwise combination of SNP loci approximately 1Mb window apart was calculated as the squared correlation coefficient (r^2) with a sample size of 98 adult female *O. volvulus* worms. SNP pairs were partitioned into bins in 1 kb intervals, and for each bin the mean r^2 was plotted against the distance between the SNP loci. Region ‘a’, ‘b’ and “c’ sections of OM1 are on 25 -26 Mb window in these figures.

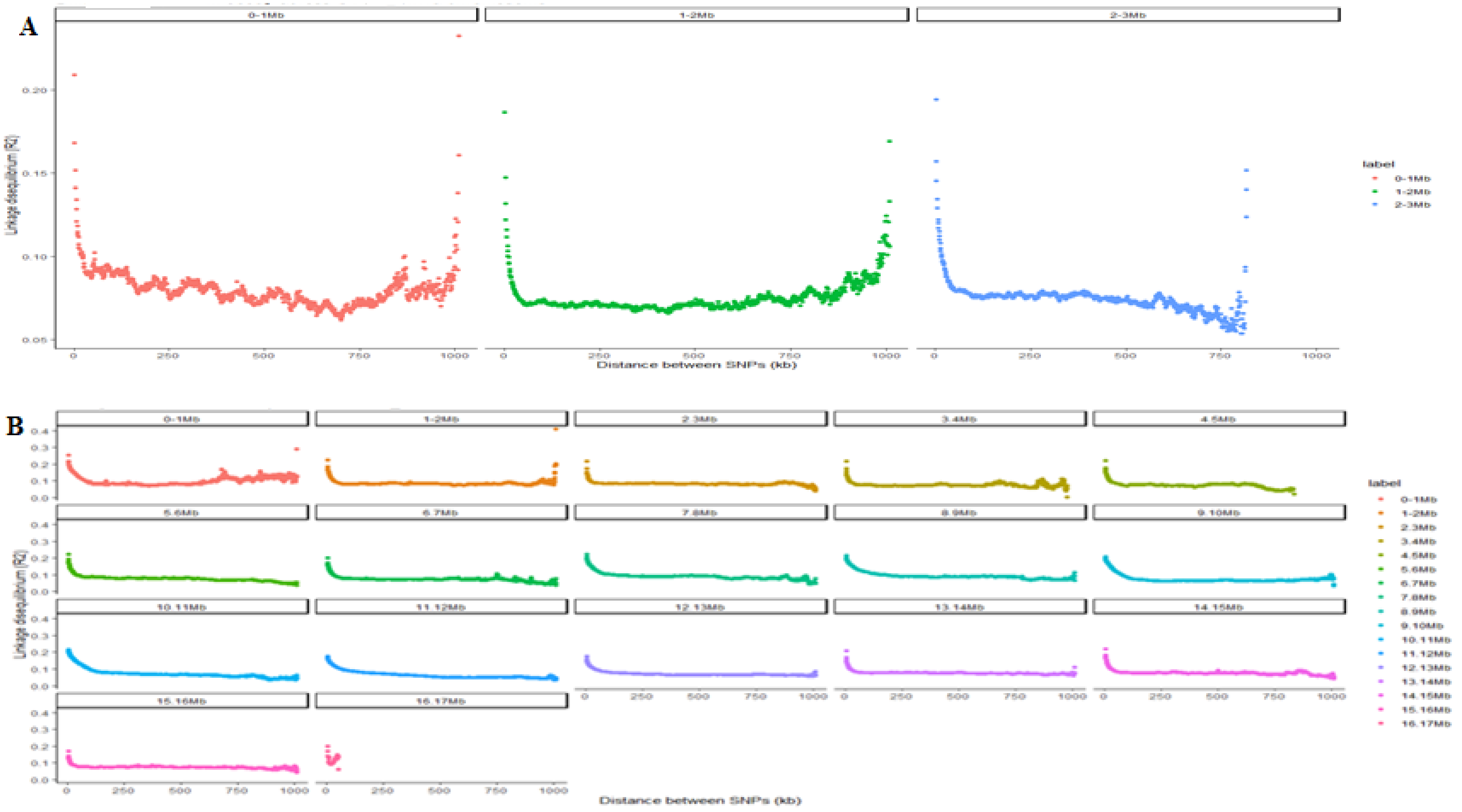


Figure 3.8b. Linkage Disequilibrium (LD) decay plots for chromosome OM1a and OM4 divided into 1 Mb window

Average LD for all pairwise combination of SNP loci approximately 1 Mb window apart was calculated as the squared correlation coefficient (r^2) with a sample size of 98 adult female *O. volvulus* worms for (A) Chromosome OM1a and (B) chromosome OM4. SNP pairs were partitioned into bins in 1 kb intervals, and for each bin the mean r^2 was plotted against the distance between the SNP loci.

3.3.3. Haploblock Structures.

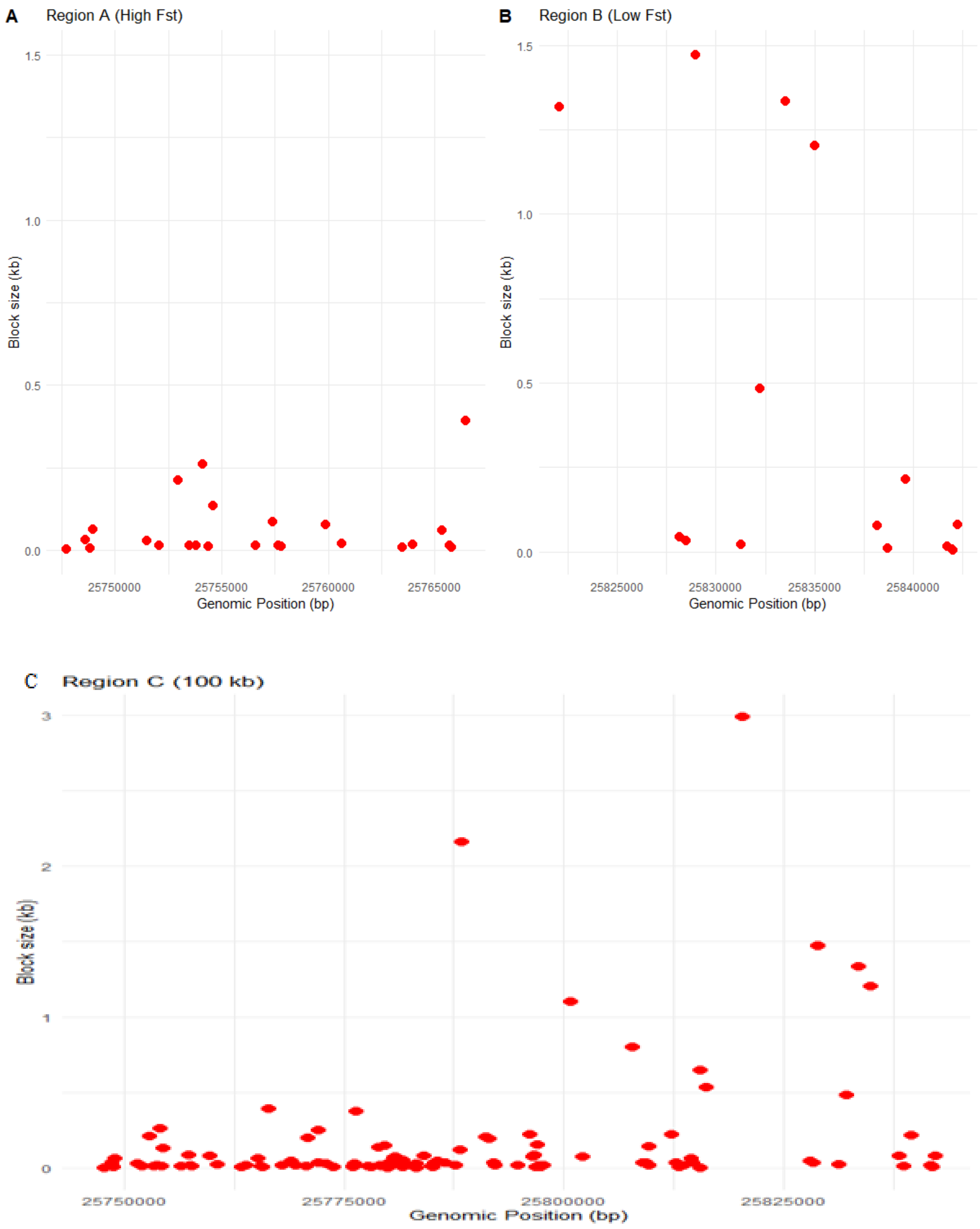


Figure 3.9. Haploblock estimate in the regions of interest.

Scatter plot of the haploblock structure against the chromosome positions within (A) region A (High F_{ST}), (B) region B (low F_{ST}) and (C) region C is shown above. Region A and B are approximately 20 kb in size and region C is approximately 100 kb in size and covers both region A and B.

3.3.3.1. Haploblock structure in regions A, B and C.

As previously discussed, a haploblock is defined in this chapter as sets of consecutive sites between which there is little or no evidence of historical recombination; that is, a site on the chromosome having LD (*D*') between adjacent SNP pairs within the confidence bound of 0.7 at least and 0.98. Based on this definition, only a small percentage of the adjacent SNP pairs made up the haploblocks in this study. The statistics of the haploblocks discussed in this section are described in table 3.5.

The size of haploblocks in region A (high F_{ST} region) and region B (low F_{ST} region) was plotted against chromosomal position to visualise the spatial pattern of LD within those regions (Figure 3.9). The figure shows that region B (the low F_{ST} region) differed from region A (the high F_{ST} region) with respect to the number, size, and clustering of the haploblocks. There were relatively few haploblocks in region B (4 haploblocks >1kb, mean = 421 bp) but they are larger compared to region A (25 haploblocks, all <1kb, mean = 63bp). There were more SNP loci in haploblocks in region B than in region A (approximately 128 versus 98), with an average of 6 SNPs per block in region B compared to 3 per block in region A (Table 3.5). Furthermore, the largest single haploblock was observed in region B (1471 bp) while the largest haploblock in region A was 392 bp. Given that region A contains the candidate "ivermectin response QTL" these data suggest that ivermectin treatment is not the principal determinant of haploblock structure in *O. volvulus.* This was confirmed when one considered the haploblock structure of the entire 100kb region. I observed haploblocks of varying length that mirrors available genetic diversity of the two regions, rich in A and poorer in B. There lengths suggest association with ivermectin resistance, but their pattern is contrast to the expectation of a selective sweep driving high F_{ST} and long-range LD (Table 3.5, Figure 3.9).

Table 3.5. Haploblock structure statistics.

Haploblock structure by region													
				Haploblock size (kb)					SNPs in haploblock				
Locus	**Length of Region (bp)**	**No of SNPs**	**No of Blocks**	**Total Block Size**	**% of Region Length in Blocks**	**Mean**	**Min**	**Max**	**Total No of SNPs**	**% of SNPs in Blocks**	**Mean**	**Min**	**Max**
A	21,368	676	25	1.509	0.006	0.063	0.002	0.392	98	14.497	4.083	2	14
B	20,623	375	16	6.316	0.021	0.421	0.002	1.471	128	34.133	8.533	2	21
C	96,004	2,699	109	18.563	0.03	0.172	0.002	2.987	651	24.120	6.028	2	54

Haploblock structure by chromosome													
				Block size (kb)					SNPs in block				
Chrom	**Length of Chrom (bp)**	**No of SNPs**	**No of Blocks**	**Total Block Size**	**% of Chrom Length in Blocks**	**Mean**	**Min**	**Max**	**Total**	**% of SNPs in Blocks**	**Mean**	**Min**	**Max**
OM1	31,161,767	422,125	18,721	6,272.490	0.020	0.360	0.002	33.44	87,533	20.897	4.720	2	463
OM4	16,048,563	233,375	10,536	2783.469	0.017	0.264	0.002	20.8	46,984	19.589	4.459	2	128

The sum of SNP loci and haploblocks, total, minimum and maximum haploblock length, percentage of the sequence covered by haploblocks, number and percentage of SNPs in haploblocks on a per chosen region and chromosome are reported in this table

3.3.3.2. Haploblock structure across OM1 and OM4.

Variation in the structure of haploblocks on the selected regions and chromosomes are shown in Table 3.3. Generally, the pattern of haploblocks on chromosome OM1 differed from chromosome OM4, primarily because it is larger (approximately 31 Mb versus 16 Mb) and had more SNP loci (422,125 versus 233,375). As a result, chromosome OM1 had more haploblocks (18,721 versus 10,536) than OM4. The percentage of the chromosome length in haploblocks was similar for both chromosomes (0.020% versus 0.017%), and those haploblocks had a larger average size in OM1 compared to OM4 (approximately 360 bp versus 264 bp). The haploblock with the longest span was larger in OM1 (33,439 bp at chromosome OM1 positions 27368457 to 27401895) compared to OM4 (20,798 bp at chromosome OM4 positions 10751001 to 10771798) (Table 3.3). Likewise, the SNP density for each chromosome varies per window size. This affects haploblock estimation since LD (*D'*) is a function of the distance between SNP loci for which it is calculated. On average, LD (*D'*) was larger where SNP density is lower. An average SNP per 1 kb of 13.95 was observed on OM1 (one SNP per 74 bp on average) and 14.95 SNPs on OM4 (that is, one SNP per 68 bp on average) (Table 3.3). The average interval between each haploblock is approximately 1,162 bp in OM1 and 1,260 bp in OM4.

The percentage of SNPs in a haploblock was similar for OM1 and OM4 (20.897% and 19.589%, respectively). On average, the number of SNP loci per haploblock was also similar in both chromosomes (approximately 5 SNPs per block). The number of SNPs within each haploblock varied and correlated positively with haploblock size (Pearson's product-moment correlation = 0.617 for OM1 and 0.73 for OM4, with p-value of $< 2.2 \times 10^{-16}$ for both). Haploblock sizes and F_{ST} values do not correlate at the chromosome level as the Pearson's product-moment correlation for both chromosomes were near 0. This is important in that F_{ST} is a measure of genetic differentiation. This lack of correlation indicates that whatever has caused the GR and SOR sub-populations to diverge genetically has not affected LD at the chromosomal scale. That is, SOR and GR have diverged by F_{ST}, but that divergence is localised to specific chromosomal regions. This means that F_{ST} and LD (which also varies in localised regions only) are not going to be correlated generally across the full length of a chromosome. This also means that it is unlikely that selection (or certainly hard selection) is involved.

To understand the spatial pattern of LD at the chromosome level in *O. volvulus*, LD was compared between linkage groups by estimating the haploblock structure of two of the major chromosomes (OM1 and OM4). The size of the haploblocks on the OM1 and OM4 chromosomes was plotted against the starting position on the chromosome to identify potential signatures of positive selection (Figure 3.10). The plot was overlayed with the F_{ST} plots for each chromosome. Looking at the figure, it is interesting to see some localised "peaks" of increased haploblock size (that is, showing "peaks" of LD). Just as the peaks of F_{ST} only "emerged" when plotted in windows, perhaps the LD "peaks" are "emerging" when plotted not as individual values but as haploblocks. With respect to the number, size, and clustering of the haploblocks, more clusters of smaller haploblocks were observed in OM1 compared to OM4. Most of the blocks are clusters of small haploblocks within the size of 2 bp to 2500 bp, although there are a few haploblocks having sizes of 10,000 bp and above on both chromosomes.

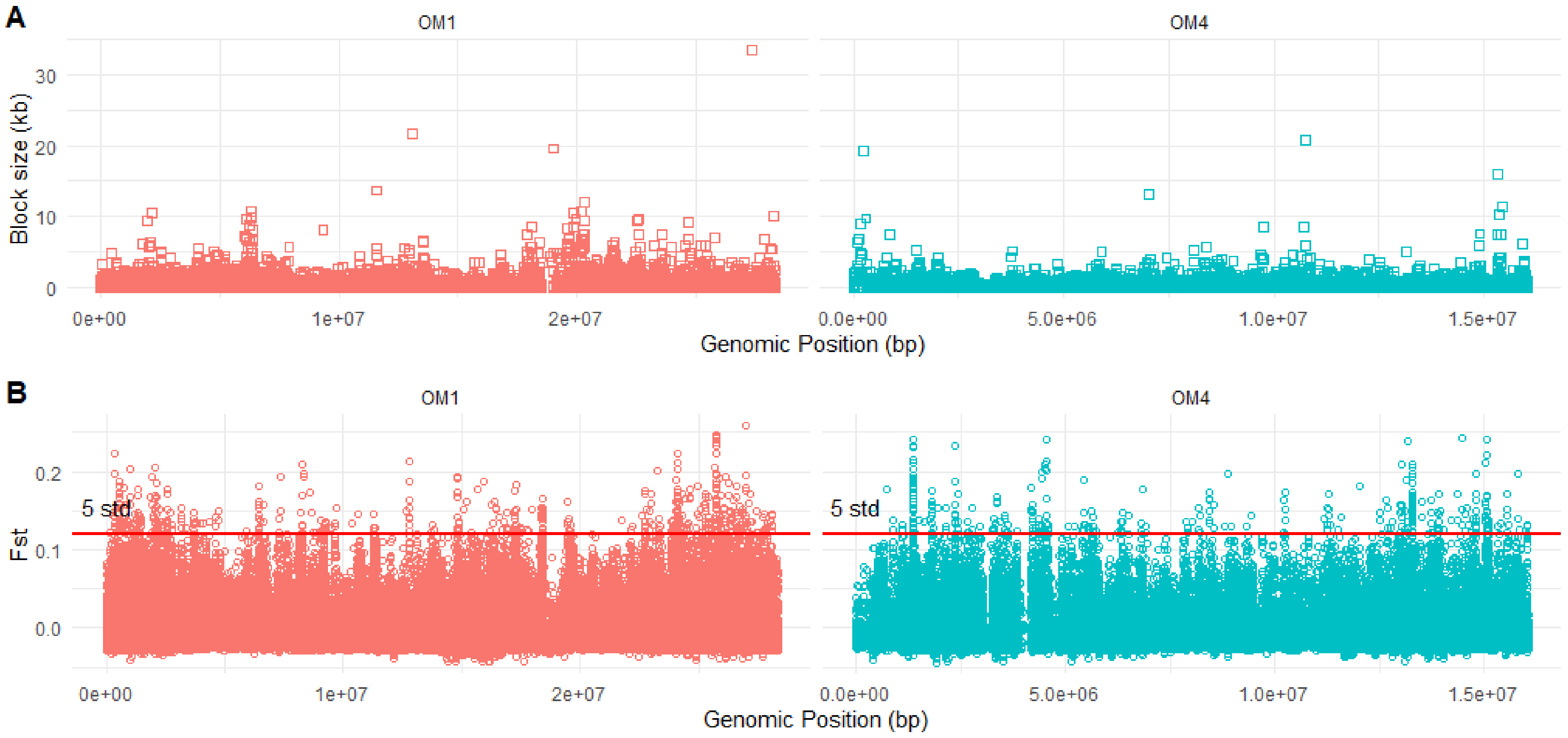


Figure 3.10. Haploblock and F_{ST} plots versus genomic position across two autosomes of the *O. volvulus*.

Plot reveals the genomic architecture of two autosomes. LD varies across the chromosome length as observed in the haploblock plot (plot A). LD blocks of >30 kb was observed in chromosome OM1 while LD blocks were not as large in OM4 (max block size of approximately 20 kb). More clusters of small blocks were observed in OM1 compared to OM4. Most of the blocks are clusters of small blocks within the size of 2 bp to 2500 bp, while fewer standalone blocks are 10,000 bp and above for both chromosomes. Localised peaks of increased haploblock sizes were observed in plot A

3.4. Discussion

This is the first study to characterize the pattern of LD and haploblock structure across the autosomes of the parasitic nematode *O. volvulus*. The major aim of this study is to understand the structure of LD and haploblock between sub-populations (GR and SOR) and across the length of two autosomes. This study has compared the variation in LD pattern and haploblock structure at several genomic scales: between two regions which are and are not thought to be associated with SOR, and between two autosomes. The major outcomes of the study are the clear understanding provided about the required LD for successful imputation in *O. volvulus*, the density, spacing, and number of useful SNP loci essential for association studies, and identifying the density of SNP loci needed for designing a SNP array for future use in developing diagnostic tools.

3.4.1. Linkage disequilibrium decay/extent.

In this study, r^2 was used to measure LD across all SNP pairs. I measured (1) LD between a local maximum of LD and SNP loci that are increasingly further away (Figures 3.3, 3.4 and 3.6) (this is an important quantity for GWAS); (2) LD between all possible pairwise without reference to their positions relative to each other or to other SNP loci (Figures 3.8 A and B, Tables 3.3 and 3.4); and (3) LD between adjacent pairs in a sliding window moving along a chromosome (Figures 3.2, 3.5, 3.7, 3.9 and 3.10).

LD decay was defined as the distance over which the average r^2 declines below 0.33 (the LD threshold for GWAS), and LD extent as the distance over which the average r^2 fell to an asymptote value below 0.05. Definitions of LD decay and LD extent vary between studies. For example, Angius et al. and Garcia-Gamez et al. used the distance over which the average LD decreases to half of its maximum value (half-length) to define LD decay in Sardinian population isolates and Spanish Churra sheep, respectively (Angius et al., 2008, García-Gámez et al., 2012), whereas Anderson et al. defined LD decay as the distance over which the average r^2 dropped below 0.8, and LD extent as the distance over which the average r^2 fell below 0.2 (Anderson et al., 2018).

Cutter et al. (2006) gave a general description of LD pattern found in *O. volvulus* populations, using six short nuclear loci examined in a broad geographic sample of wild isolates of the gonochoristic *C. remanei*. They observed that LD declines significantly over just a few hundred base pairs at a rate suggesting that linkage equilibrium will be reached at distances of 1–2 kb (mean r^2 of 0.208 across the loci at an average distance of

1781.1 bp). My results, as seen in Figures 3.3 and 3.4A show that LD surrounding a candidate ivermectin response-QTL decayed from the starting value to r^2 values of 0.33 at an approximate distance of 1404 bp and extended to an asymptote level of 0.05 within the approximate distance of 9 – 10 kb. This contrasted strongly with LD decay surrounding a local maximum of LD that was not associated with an ivermectin response QTL (Figure 3.4B) in which r^2 decayed below 0.33 within 50bp of the local maximum, while still reaching the asymptote of approximately 0.05 also within 9 – 10 kb. Thus, this analysis of LD decay around local LD maxima is consistent with a relatively weak but none the less detectable LD signal that is associated with ivermectin response (or may also be related to differences in the number of SNPs available). The relatively weak nature of this signal requires, however, that SNPs are genotyped at intervals of no more than approximately 1.4kb (in this instance). This is similar to the observed average r^2 across 6 short nuclear loci examined in a broad geographic sample of wild isolates of the gonochoristic *C. remanei* by Cutter et al. (2006). A similar rapidly decaying LD was observed in *Drosophila melanogaster* and maize (Long et al., 1998, Remington et al., 2001, Tenaillon et al., 2001). Whereas LD decay in humans and in *Arabidopsis thaliana* is as large as 50–60 kb (Nordborg et al., 2005, Reich et al., 2001).

There is variation in the average measure of the relationship between LD and the distance between SNP loci across all pairwise combinations within the selected regions and across the chromosomes studied in this chapter. These measures of LD are not related to the LD decay and extent discussed in reference to Figures 3.3, 3.4 and 3.6 but are of interest as long-range measures of variations in recombination between chromosomes or large segments of chromosomes. That is, they are useful to get the general sense of chromosome level differences of LD. LD between SNP loci across chromosome segments is the basis for haploblock definition that relies heavily on subsequent GWAS. Recombination also constrains gene flow between population which greatly affects F_{ST} values. The difference in the average measure of the relationship between LD and the distance between SNPs at various 1 Mb segments of the chromosome length (Figures 3.8A and 3.8B and appendix Table 3.1) could be attributed to recombination rates varying between and within chromosomes, genetic drift, and demography. Differences in LD between chromosomes have already been reported in Holstein cattle (Qanbari et al., 2010).

3.4.2. Haploblock estimation and correlation with selection.

The major aim of studying the haploblock structure of *O. volvulus* was to identify the pattern of LD variation along the linear sequence of a chromosome. Based on the definition of haploblock in this study, a region is described as a series of small, adjacent, clustered blocks or these are merged into a longer single block. Defining haploblocks is essentially an extension of the LD sliding window that groups adjacent windows of elevated LD together into blocks. For LD between adjacent SNP pairs, Figure 3.2 shows the sliding window LD plot between adjacent SNP loci on a 100 kb region on chromosome OM1. From the figure, a small interval or peak of elevated LD was detected around the first 20 kb which matches with a region of high F_{ST} as shown in Figure 3.2B and 3.5A&C although small intervals of elevated LD existed across the whole region. This could suggest soft selection, because a small window of moderate elevation of LD involving a small number of adjacent SNPs is expected in soft selection. However, the relationship between F_{ST} and LD does not just depend on selection. It could also be because of lower recombination in that region which favours reduced gene flow and, hence, higher $F_{ST.}$ F_{ST} is conditioned by within population diversities and gene flow between the two populations compared. And, LD could increase because of diversity loss, for example, selection, or reduced recombination (functional constraint or genomic constraint). Clustering of shorter haploblocks might indicate selection. I compared two regions, one (region A) which is associated with SOR based on high F_{ST} between GR and SOR worms and the other (region B) which is not (Figure 3.5). What emerges is a view that where there is elevation of F_{ST} there is also modest elevation of LD, but the converse is not necessarily true, that is, there are many regions of elevated LD that do not correspond with elevation of F_{ST} and, in fact, overall, there is poor correlation between the two. In other words, the existence of haploblocks that are not associated with elevated F_{ST} does not falsify the hypothesis that regions of elevated ivermectin associated F_{ST} will be associated with a haploblock. The distinct clustering of moderately elevated LD haploblocks in region A aligns very well with the F_{ST} plot. Considering the clustered pattern of the short haploblocks in region A, it is possible that LD is elevated over a much longer region in A than in B even though the individual haploblocks in B are longer (Figure 3.9 A&B). This corroborated the result shown in Figure 3.5 on the LD between pairs of adjacent SNP loci within region A and B, where I observed a region of elevated LD extending over 10 kb in region A, consistent with selection and supported by the LD decay rate described in Figure 3.4.

Higher LD and resulting longer haploblocks are expected in regions that have undergone selection or are presently undergoing selection in the population. The haploblock with the longest span observed in region B (a low F_{ST} region) was approximately 4 times than the longest haploblock observed in region A (a high F_{ST} region) (Table 3.3; Figure 3.9). When considering the larger region of the chromosome in which region A and B are located, region C, there are six (6) haploblocks and they are not clustered. The F_{ST} peak in region A is mirrored by a moderate (2 – 3-fold above average) elevation in LD in a haploblock, and there is a much larger region of elevated LD between regions A and B that does not correspond to a peak in F_{ST}, and evidence for a very large region of elevated LD around position 25,820,000 just to the left of region B. Thus, the elevation of LD in region B is because it is on the shoulder of an extensive region of LD (Figure 3.9) that is not mirrored by elevated F_{ST}. This implies that SOR selection has left a weak LD signature consistent with soft selection (in Region A) but that there has been much stronger selection that is not related to SOR further to the right of Region A. The nature of this selection is not known, but a reasonable candidate might be the host switch that gave rise to *O. volvulus* or there could be another genomic artifact in observed signal, for example, the region may be poorly resolved.

SNP density has a strong impact on the ability to detect small haploblocks. This could contribute to the difference between the number of blocks found in region B (low F_{ST} region) compared to A (high F_{ST} region) - there was a 2-fold difference in the percentage of SNP loci in haploblocks and the 2-fold difference in mean SNP loci per haploblock in region B compared to region A. There was also a 2-fold difference in SNP locus density between region A and B (Region A= 676 SNP loci; B = 375 SNP loci). The definition of a haploblock in this study (pairs of SNPs having LD values between *D'* of 0.7 – 0.98) combined with SNP density in region A and B (region B had a SNP locus density half that of region A) could account for such apparent difference in the haploblock structure between regions A and B, and certainly between OM1 and OM4 (where SNP density is similar, as are percentage of chromosome in haploblocks and percentage of SNP loci in haploblocks). Haploblock coverage (the proportion of sequence that is contained in haploblocks) can potentially be improved by using higher SNP densities. Both simulations and biological data show that an eight-fold greater SNP density improved the proportion of sequence that is contained in haploblocks more than twice, most of which comes from the identification of smaller blocks that were missed with sparser SNP density (Wall and Pritchard, 2003b).

Less SNP could result from less diversity, and, thus, increased linkage and increased haploblock size. On the average, the distance between adjacent SNP loci is greater in region B than in A resulting into the less likelihood that any two adjacent SNP loci are sufficiently close together to be in LD, and thus to form a haploblock. The numbers support it up to this point (676 vs 375 SNP loci in A vs B; 25 vs 16 blocks) but the situation changes when one looks at the percentage of the region in haploblocks (0.006 vs 0.021 or 3-fold less in A vs B), which translates to a 7¬fold more base pairs of sequence in haploblocks in B than A and more than 2¬fold more SNP loci in haploblocks in B than in A. All those numbers that has to do with haploblock size are the opposite of what one expects in a region in which SNP locus density is lower. This implies that the difference in SNP locus density may not account for the difference in haplotype structure between the two regions (that is, SNP density does not influence haplotype structure in the regions studied). Another explanation could be because of the problem of sample size. Recombination is stochastic, so the ability to measure LD accurately is dependent on sample size. This was evident with the region with higher SNP density such that low sample size (as seen here) resulted in variation in LD over short distances and shorter haploblocks.

The genome-wide pattern of LD/haploblocks showed that some regions that differentiate SOR adults' female worms from GR correlated well with regions of elevated LD (consisting of a cluster of fragmented small haploblocks) while others do not, suggesting a pattern of soft selection for ivermectin response that varies between GWAS-QTL loci. There are many other influences on LD in addition to ivermectin response, including drift, migration/admixture (which can both distort F_{ST} and LD); or selection at closely related locus; or a loss of diversity at region B could be driving the pattern at region A. Another possibility is that the F_{ST} analysis is impacted by sample size (the same is true, possibly even more) for estimates of LD because recombination is itself stochastic. some of the high F_{ST} values will be stochastic, rather than because of selection for SOR. One expectation is that high F_{ST} values that are due to selective forces, rather than sample size, will be associated with regions of high LD or be within haploblocks. Sweep will distort LD (Thomson 1977) but the relationship between F_{ST} and LD involves gene flow, population diversity and recombination (Slatkin and WIehe, 1998). In addition, LD distortion depends on the age of the sweep. To that respect, the perspective of using haplotypes of SNPs for F_{ST} computations could help reducing noise associated with SNP-based analysis (Charlesworth, 1997).

During the domestication of cattle by humans, there is likelihood for *O. volvulus* host switch from cattle to humans. In domesticated cattle breeds like Holsteins, N_e estimate is small compared to *O. volvulus* (that is, lower recombination and higher LD). Similar estimates were expected in *O. volvulus* population, but the converse is true: higher recombination, lower LD between SNP loci and less of *O. volvulus* genome in LD were observed. For example, the percentage of the chromosomal length contained in haploblocks was approximately 0.02 % and there were approximately 5 SNP loci per haploblock on average (Table 3.4). The genomic distribution of, and proportion of the genome covered by, haploblocks in this *O. volvulus* population is lower than observed in some other species such as cattle (Kim and Kirkpatrick, 2009, Qanbari et al., 2010), as expected according to the lower LD between SNP loci observed. The situation is also different in sheep: the SheepHapMap project identified an overall limited genome coverage in haploblocks for domestic breeds, with Churra having the lowest coverage (0.8%) with 88% of the blocks had 2 SNPs (Qanbari et al., 2010) and the wild Soay sheep showing large genome coverage (21.84%) (Archibald et al., 2010).

3.4.3. Linkage Disequilibrium and evaluation of effective population size.

There are two general contributors to variation in LD: one is recent events (selection) and changes in population size (population bottleneck and admixture). The final value of LD at any point in the genome is a trade-off between these two. The measurement of LD in natural populations has been utilized to estimate the effective population size (N_e) (Deng et al., 2019, Service et al., 2006, Sved, 1971, Tenesa et al., 2007). N_e can be described as the number of individuals in an idealized population with random mating and no selection that would lead to the same rate of inbreeding as observed in the real population (Wright, 1931). One important relationship between LD and N_e is that LD between SNPs farther apart is a reflection of a more recent change in N_e than LD between SNPs closer together (Hayes et al. 2003). A large N_e means either that the census population size is very large and/or that there have been many generations of recombination: the chromosome segments that are identical by descent are small, and so LD extends for only a short distance (reviewed in Ardlie et al. 2002). Higher recombination and lower LD between SNP loci observed in *O. volvulus* populations is consistent with a large effective population size and a large census population size in the parasite populations. This result was corroborated by Crawford et al. (2019) in their study using whole mitochondrial genomes of approximately 150 parasites from West Africa; they indicated that historical population size in *O. volvulus* is likely very large (10^5 or 10^6). Crawford et al.'s analysis is consistent with the hypothesis that there was a population bottleneck in the distant past followed by subsequent expansion. Similar values for N_e were also estimated on the basis of the nuclear data from Ghana (S. Hedtke, pers. comm.).

Realistically, the N_e of a population can change over time (Wright, 1931). For instance, in *Bos taurus* cattle N_e was large before domestication (>50,000), declined to 1,000–2,000 after domestication and, in many breeds, declined to approximately 100 after breed formation. This was coupled with the very strong artificial selection for production traits in the livestock. However, the long-range LD only applies in similar breeds (MacEachern et al., 2009). This N_e history in cattle is similar to that experienced by dogs (Sutter et al., 2004) but is the opposite of that experienced by humans (reviewed in Ardlie et al. 2002). The European human N_e was only approximately 3,000 but then increased enormously in the last 10,000 years. Consequently, humans have similar LD to cattle at short distances but almost no LD at long distances (Ardlie et al., 2002, Tenesa et al., 2007). Given the results from Crawford et al, there was a huge bottleneck in *O. volvulus* approximately 10,000 years ago, probably during the host switch from cattle to humans (similar results have been observed but not reported for the nuclear data for Ghana; S.Hedtke pers comm).

3.4.4. Linkage Disequilibrium and GWAS.

Generally, r^2-values above 0.33 indicate LD that is sufficiently strong enough for association studies that use a subsample of the data (Ardlie et al., 2002). This threshold was based on the attempt to interpret r^2 in terms of power to detect an association (Kruglyak, 1999). Typically, sample size is a limiting factor in association studies (Spencer et al., 2009), but increasing sample size to compensate for weak LD between a locus and the susceptibility QTLs is impractical in *O. volvulus*, because of limited ability to acquire phenotyped samples, inconsistent quality of the gDNA extractions, cost of genomic sequencing, and resulting missing data in the sequenced worms (Doyle and Cotton, 2019, Hedtke et al., 2019)). Values of $r^2 > 0.33$ limit the required increase in sample size to no more than threefold and was therefore considered to be the minimum useful LD value for GWAS. Application of this rule of thumb to statistical averages in my data based on the physical distance between the adjacent SNP loci with $r^2 > 0.33$ would imply it is not possible to carry out GWAS successfully unless markers are spaced at an average physical distance of 56 bp and 57 bp between adjacent SNP loci on chromosomes OM1 and OM4 respectively.

Furthermore, because the values of LD vary across the chromosome length, the criteria for SNP density varies as well. For example, based on the physical distance between the adjacent SNP loci where $r^2 > 0.33$ in region A (high F_{ST} region), an average physical distance of 28 bp between adjacent SNP loci would be required for GWAS, while in region B (low F_{ST} region), this value is 63 bp. Similarly, in the 100 kb region that covered both A and B in this study (region C), an average physical distance of 42 bp between adjacent SNP loci would be required for GWAS. These are evident in Figures 3.2 and 3.5 (points above the horizontal line).

The extent of LD serves to assess the number of markers required to associate genetic variation with economically important traits (García-Gámez et al., 2012). A population with extensive LD will require lower marker density; in contrast, if LD extends for only a short distance, denser SNPs would be needed for GWAS to detect or increase the power of association. Figures 3.3 and 3.4A suggest that SNPs can be further apart in region A than in B because LD persists further. This means that lower SNP density is required in region A than in B because LD surrounding the candidate QTL that is under selection decays more slowly in A than in B. Based on the extent of LD up to r^2 value of 0.33 (which is at approximate distance of 1.5 kb), a minimum SNP density of 20,775 and 10,699 are needed in OM1 and OM4, respectively, to confidently detect the association of a SNP with a trait of interest in *O. volvulus*. In other words, If and only if the candidate ivermectin-QTL studied were typical of all ivermectin-QTLs, SNP loci would have to be no more than 1.5kb apart in order to detect selection at an ivermectin-QTL. This is similar to analysis done by McKay et al. in cattle, in which they showed that at a physical distance of 100 kb separating flanking SNP loci, the average r^2 was 0.15 to 0.2; considering a bovine genome length of 2.87 Gb, they concluded that 28,700 fully informative markers would be needed to saturate the cattle genome at an average resolution of 100 kb (McKay et al., 2007). Alternatively, genotype imputation from a densely genotyped reference panel could help to improve the density of SNPs in situations of missing or ungenotyped markers. This is discussed elaborately in Chapter 4.

The estimation of LD reported in this chapter can also help to assess the utility of a SNP array genotype chip to address fine-mapping studies in *O. volvulus*.

3.5.Conclusion

This is the first study to characterize the pattern of chromosomal LD and haploblock distribution in the adult females of the parasitic nematode - *O. volvulus*. I have been able to describe LD pattern within specific high and low F_{ST} regions on the OM1 chromosome of the *O. volvulus* genome across worms from Ghana phenotyped for response to ivermectin. I have further extended this analysis across two autosomes.

This study has implications for using LD as a tool to study population history, in the design of a SNP array, for development of diagnostic tools, to allow for genotype imputation and for enhancing GWAS in *O. volvulus* and helminths at large. The major highlights of this chapter are:

1. Across the regions and chromosomes studied, LD decayed within few base pairs and extended to an asymptote level of r^2 of 0.05 within the approximate distance of 7 – 9 kb.
2. There is sufficient LD for genomic imputation (high number of adjacent SNP loci with $r^2 \geq 0.33$) which is the focus of the next chapter. Based on the LD data, the required SNP density needed to confidently carry out genotype imputation is a SNP locus every 56 or 57 bp. That is, markers must be very closely spaced (at least at an average physical distance of 57 bp between adjacent SNP loci across the genome).
3. GWAS is dependent on LD decay around local maxima of LD (and haploblocks) that are indicative of selection. Based on the extent of LD up to r^2 value of 0.33 (which is at approximate distance of 1.5 kb), a minimum SNP density of 20,775 and 10,699 are needed in OM1 and OM4, respectively, to confidently detect the association of a SNP with a trait of interest in *O. volvulus.* A minimum of 64,668 SNP loci are needed at the genome level.
4. Soft selection is driving SOR in *O. volvulus* (although, subjected to further analysis), which in turn left a weak LD signature in the genome characterised by clusters of small fragmented haploblocks of low to moderately elevated LD that correlate with peaks of F_{ST}.

5. There is clear evidence of much stronger selection that are not related to SOR. The nature of this selection is not known, but a reasonable candidate might be the host switch that gave rise to *O. volvulus.*

6. The LD results are consistent with a large effective population size and a large census population size in the parasite population as estimated from previous studies.

Chapter Four

Genomic imputation

Introduction

With the emergence of the genomic era, there is an exceptional amount of genomic information being generated routinely (Isik et al., 2017). Inevitably, some fraction of the genotype data will be missing from virtually every genotyped dataset. For single-marker methods, such as association genetics, removing the individual or locus records in which these missing values occur could be an option, provided that the overall level of missing data is relatively low for each locus. Analytical methods that consider more than one locus at a time are often much more sensitive to missing data, and it may not be economical to drop individuals that are missing data at any one of the loci being analysed, unless they are few (Isik et al., 2017). In the extreme case, in which all marker loci in the genome are analysed at once, the approach of dropping individuals with missing genotype data can mean discarding data which may otherwise be informative about the phenotype of interest in order to remove a small proportion of missing values (Neale, 2010, Illumina, 2019). A valuable alternative is genomic imputation, which maximizes the amount of information available for analysing genotype-phenotype correlations.

Genotype imputation refers to the estimation of base calls to replace missing observations in a data set (Isik et al. 2017). It is a cost-effective method for statistically predicting un-typed loci not directly sequenced in a sample of individuals based on a densely genotyped reference panel of haplotypes (Neale, 2010) and can be almost as accurate as directly sequencing the genotypes. Imputation methods estimate haplotypes based on shared genotypes between partially genotyped individuals and the reference panel and use this information to infer missing alleles (Marchini and Howie, 2010). That is, it relies on a reference database of fully sequenced genomes to predict missing variant calls in a sample of individuals. The approach consists of first reconstructing haplotypes for the samples of interest (samples of interest are described as a ‘target population’ in this chapter) using the haplotypes from the reference set (haplotype phasing), and then estimating genotypes (Neale 2010). This process can increase the overall genome coverage of any genotyped dataset by increasing the number of testable single nucleotide variants across the entire genome and can improve fine mapping of a targeted region of interest (Illumina, 2019).

Imputation methods exploit linkage disequilibrium (LD) among SNP loci. LD is a powerful asset in imputing missing genotypes with 90% or greater accuracy (Huang et al., 2012). Imputation essentially relies on knowing or inferring the haplotype phase of individuals and then using LD information from nearby markers to replace missing genotypes. The structure of LD between adjacent SNP loci described in chapter three, section 3.3.2.3 and in figure 3.7 of this thesis shows that in the *O. volvulus* genome, there is sufficient LD between adjacent SNP loci to carry out genotype imputation because a sizeable percentage (approximately 26.48% and 22.29%) of the adjacent SNP loci across the autosomal chromosome OM1 and OM4, respectively, had LD above 0.33 (the LD threshold for imputation). Although genotype imputation has not been tested in helminths, it could be a powerful tool for minimizing costs associated with genetic-based screening for sub-optimal response in *O. volvulus* or for drug resistance in other helminths.

Imputation is new to filarial nematode population genetic studies, and reference panel (that is, densely genotyped datasets) have not been developed that could be used to impute missing genotypes in these animals. In other organisms where genotype associations are ordinarily applied, reference panels for imputation have been developed. For example, the human genome project uses the samples from the Human Genome Diversity Project the HapMap Consortium, and the 1000 Genomes Project (1000G) as reference panels (Cavalli-Sforza, 2005; International HapMap 3 Consortium, 2010; Sudmant et al., 2015); in the sheep genome project; 5K, 50K, and HD panels are routinely used (Ventura et al., 2016), the cattle genome projects generally uses the animals from the **1000 Bull Genomes Project (Run 6.0)**, **the BovineSNP50 (SNP50) BeadChip** (Illumina, San Diego, CA), **and the BovineHD (Illumina, San Diego, CA) array** as reference panels (Matukumalli et al., 2009; Wiggans et al., 2012; Daetwyler et al., 2014).

The Grant Lab at La Trobe University currently has the world's largest *O. volvulus* whole genome sequence database. I utilized this resource to test the benefits and limitations of genomic imputation in these worms. Densely genotyped *O. volvulus* were used as the reference panel for imputing missing genotypes in low and uneven-coverage sequences from different worm populations. The overall aim of the project, and of the Grant Lab, is to develop diagnostic tools for identifying SOR genotypes from hundreds of *O. volvulus* microfilariae collected from people. Imputation may be a practical solution to reduce the costs associated with genotyping the microfilariae because each microfilaria has little DNA, and because sequencing hundreds at high depth would be impractically expensive, while the more economical low-depth whole genome sequences have correspondingly low genome coverage. Thus, the conceptual feasibility and accuracy of imputation were explored in this chapter, with regard to (1) identifying the appropriate reference panel to use by comparing two sets of reference panels: (a) from a diverse population (called 'global' in this chapter) and (b) from the same population as the target population, and (2) testing the success of imputation in improving the power of association of variants with drug resistance.

Materials and Methods

4.2.1. Study Data.

The worm samples used in this study consist of adult female *O. volvulus* from East Africa, West Africa, and South America (total = 192). Figure 4.1 shows the countries from which the worms were obtained and their assignment into a reference panel (that is, the worm sequences containing a dense proportion of SNP loci used to impute missing genotypes) and a target population (worm sequences with missing genotypes or genotyped at lower marker density).

There were two imputation experiments. In one, a “global” reference panel made up of the worms obtained from NCBI’s GenBank database, study accession number SRP066374 from a study by Choi et al. (2016) was used to impute data for population of worms from W. Africa (Grant lab; sequenced at lower and inconsistent depth relative to the global reference), and from Cameroon (reduced representation sequencing, also from Grant lab). In the second experiment, population specific reference panels were selected from Grant lab sequenced worms from W. Africa and Cameroon, where those worms with higher quality and higher depth whole genome data were selected as reference panels and the remaining lower quality or reduced representation worms were the target populations. The aim was to compare different strategies for choosing a reference panel and to test the feasibility of using genomic imputation as an alternative to high depth sequencing for genotyping *O. volvulus*.

For the GWAS study, a subset of the W. Africa and Cameroon targets populations had been phenotyped for sub-optimal response and thus could be used for association analyses. Worms from W. Africa were qualitatively phenotyped based on embryograms from nodulectomies approximately 90 days after ivermectin treatment and categorized by response to ivermectin into good responders (GR; n = 33) and sub-optimal responders (SOR; n = 26) (Figure 4.1 and as described in chapter two Methods section). The worms from Cameroon were phenotyped based on quantitative counts (of embryonic developmental stages including number of oocytes, oocytes in rachis, morula, coiled microfilariae, and stretched microfilariae from embryograms taken from nodulectomies approximately 80 days after ivermectin ingestion). Worms with higher counts of the embryonic stages were categorized as suboptimal responders (SOR, n = 32) and the ones with no counts were categorized as good responders (GR, n = 33) (Figure 4.1).

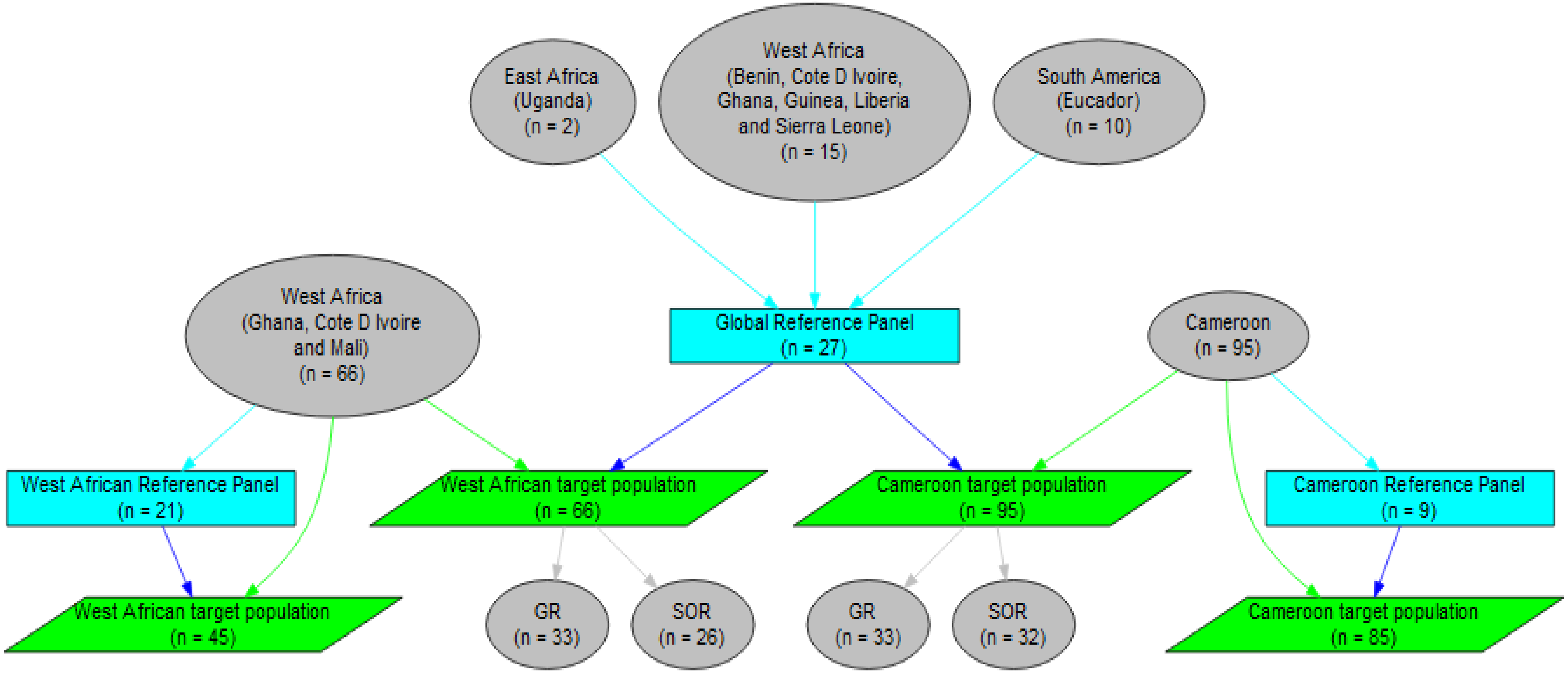


Figure 4.1. Flowchart showing countries where samples were obtained, and the distribution of these samples into reference panels, target population, and phenotypes.

Countries from which the samples were sourced: the big circles in grey colours, their phenotypes: small circles in grey colours, including good responders (GR) and sub-optimal responders (SOR)). The cyan arrows point to the reference panels (rectangles in cyan colour) while the green arrows point to the target populations (parallelograms in green colours). Finally, the blue arrows point to the reference panels used in imputing each target populations

4.2.2. Preparation of the reference panel and target populations.

Variant call files (VCF) based on the sequence data described above for the worms in the global reference panel, W. African reference panel and target populations, and from the Cameroon target population, were received from S. Hedtke (Hedtke et al., in prep), while the Cameroon reference panels were received from the lab. The global reference contained all the worms from the Choi et al. study (Choi et al., 2016).

Vcftools v1.1.13 (Danecek et al., 2011) was used to further filter all the variant call files. Only biallelic SNP loci were retained. Sites with any missing data were filtered out of all the reference panels. The global and the W. African reference panels retained only SNP loci with a minimum depth value ≥ 20, while the Cameroon reference panel had a minimum mean depth value ≥ 5. For the W. Africa target population, variant sites were filtered requiring a minor allele count (number of times the allele appears over all individuals at a particular site) ≥ 3, a minimum mean depth value ≥ 5 (over all worms in that particular data set) and <50% missingness (in order to reduce genotype error rates; also knowing that imputation accuracy improves as the depth of coverage and the minor allele frequency increase). For the Cameroon target population, individual worms with more than 50% missing data and sites with more than 20% missing data were removed. The W. African 1 (N = 66) target population are all the whole genome sequences with uneven coverage obtained from Grant Lab while the W. African 2 (N = 45) target population are a subset of the whole genome sequences with uneven coverage obtained from Grant Lab. The Cameroon 1 (N = 95) target population is all the single adult worm reduced representation sequences obtained from Grant Lab while the Cameroon 2 (N = 86) target population is a subset of single adult worm reduced representation sequences obtained from Grant Lab.

The software *beagle* (Browning and Browning, 2007) was chosen for imputation and associated analyses because it has been found to perform well with high accuracy and is more robust under various conditions as seen in previous studies (Das et al., 2016, Ma et al., 2013, Marchini and Howie, 2010). According to Browning and Browning (2007), proper imputation requires that the reference and target variant files have overlapping SNPs. A *beagle* filtering step was performed using *conform-gt* as described by (Browning and Browning, 2016) to ensure this. As described earlier, the reference panels, which were sequenced at greater depth, were used to predict missing or untyped genotypes, while target populations were the worm population where missing/untyped data were to be imputed. Using the position field, two allele records (one record each from the reference and target variant files) were matched only if the alleles in the target VCF record or in the strand-flipped target VCF record were a subset of the alleles in the reference VCF record. Positions where alleles were called in the target population that were not found in the reference panel were removed from downstream analyses.

4.2.3. Phasing of the reference panels.

All the reference panels were phased using the program *beagle* version 4.1 with default parameters (Browning and Browning, 2007). Due to small sample size, phasing was done only on the reference panels, and it ran very fast. Pre-phasing of the target population was not done to avoid introducing imputation error due to haplotype uncertainty, which can particularly occur when imputing rare variants (MAF < 0.01) (Howie et al., 2012). Also, because of the small sample size, not having pre-imputation phasing for the target population had no implication on the imputation run time. Phasing is the construction of haplotypes from unphased diploid data, that is, converting those diploid data into what is in effect two sets of haploid data for each worm. It is a critical step given the haplotypes are the basis of imputation.

4.2.4. Imputation.

Imputation of the target population was done using the default parameters in the program *beagle* version 4.1 (Browning and Browning, 2007) for two autosomal chromosomes used as test cases throughout this thesis: OM1 and OM4. *Beagle* (Browning and Browning, 2007) uses a hidden Markov model to predict the missing genotypes and linear interpolation to impute ungenotyped variants.

A large effective population size was assumed (N_e= 1,000,000) based on the estimation made from observed relatively low LD between markers in these chromosomes (average (SD) r^2 between adjacent SNP loci are approximately 0.25 (0.25) and 0.21 (0.31) for chromosome OM1 and OM4 respectively; chapter three), as well as estimates from previous studies (Crawford et al., 2019; Hedtke et al., in prep). The target populations were divided into single chromosomes, and imputation was performed for each chromosome separately. The resulting variant files were merged using the *perl* script with *vcf-concat* (Danecek et al., 2011) and used in downstream analysis.

4.2.5. Validation.

I assessed the accuracy of the imputed genotypes at the marker level using the allelic R-squared (AR2) and dosage R-squared (DR2) metrics, and at the study level using the allele-frequency correlation, as described by (Browning and Browning, 2009). Allelic R-squared (AR2) is described as the squared correlation between the allele dosage (that is, how many copies of the alternative allele are observed at a locus) of the most likely imputed genotype in the target population and the allele dosage of the true genotype in the reference panel (Browning and Browning, 2009). Dosage R-squared (DR2) is defined as the estimated squared correlation between the estimated allele dose and the true allele dose (Browning et al., 2018). Both estimates, AR2 and DR2, are good measures of imputation accuracy and either of them can be used for identifying or excluding markers with poor imputation accuracy prior to downstream analysis and interpretation (Browning and Browning, 2009, Gilly et al., 2019). A filter of AR2 >0.80 is usually encouraged - a filter that is too low could introduce errors into the imputed data (H. Daetwyler pers. comm.).

The Wilcoxon signed-rank test was used to compare the accuracy of estimated allele frequencies between reference panels and their corresponding post imputation population (that is: global/W. Africa, W. Africa/W. Africa, global/Cameroon, and Cameroon/Cameroon), according to Browning and Browning (2009). For each imputed marker *v*, if *Pv* is the absolute allele-frequency error using reference panel 1 (for example, global reference panel) and *Qv* is the absolute allele-frequency error using reference panel 2 (for example, W. Africa reference panel), the null hypothesis of the Wilcoxon signed-rank test is that the median of $Pv - Qv$ equals 0. Rejecting the null hypothesis implies that there are differences in accuracy of the estimated sample allele frequencies derived from the two reference panels.

To detect the association between the accuracy of imputation and allele frequency, the alleles were divided into 10 bins, according to their frequency, with an increment of 0.1. The accuracy of imputation (AR2) was calculated for each bin to test the efficiency of imputation for alleles at different frequencies. The correlation between the allele frequency of the reference and the post-imputation target population for each reference/target combinations was calculated to assess whether the distribution of variants after imputation was altered significantly or not. The allele frequency spectra of the post-imputation variants above or below the accuracy cut-off were compared to see if there was bias in imputation accuracy across a given chromosome.

4.2.6. Association tests.

Association tests were carried out using *plink2*v.1.90b3 standard case/control association analysis (Purcell et al., 2007) on a subset of the target W. African (n = 59) and target Cameroon populations (n = 65), which were phenotyped as described in section 4.2.1 and Figure 4.1 above. GWAS was performed to investigate the effect of imputed genotype data on the power to detect ivermectin response associations by comparing p-values computed with true genotype data (that is, the genotype from the experimentally measured population - from Illumina data) with p-values computed with imputed data. Quality control was performed on the imputed variant sites using *vcftools* v1.1.13 (Danecek et al., 2011), which involved extracting variant sites with AR2 ≥0.8 from the imputed variants, which were then used for the association studies. Case/control association test was done here on a subset of the W. Africa and Cameroon targets populations phenotyped for sub-optimal response.

To account for population stratification, multidimensional scaling (MDS) analysis was performed in *plink2* to perform a cluster analysis that pairs individuals on the basis of genetic identity with Euclidean distance.

Manhattan plots of p-values were created using the *qqman* package in *R Studio* (R Development Core Team, 2013, Turner et al., 2013). Other figures and tables were created in *RStudio* (RStudio, Boston, Massachusetts, USA) using the *ggplot2* package (Wickham, 2016) and Microsoft Excel (Microsoft Corporation, Redman, Washington, USA).

Results

4.3.1. Reference and target variant filtering.

The quality of imputation is dependent on both on the quality of the reference panel and quality of the target population. The counts of variant sites in the reference panels, the counts, and the percentage (in brackets) of variant sites in the target populations before and after running *beagle* (*conform-gt*) filtering steps, and after imputation are provided in Table 4.1 for the two autosomal chromosomes tested here (OM1 and OM4). The variant sites in the global panel (N = 21; whole genome sequences obtained from Choi et al. (2016) study) were used to impute the W. African (N = 66) (whole genome sequences with uneven coverage obtained from Grant Lab; average depth of coverage = <20) and Cameroon (N = 95) target populations (reduced representation sequences obtained from Grant Lab; average depth of coverage = <5).

After filtering the reference panels for minimum depth and quality with *vcftools*, as expected, the global reference panel had more variable sites (569,363) compared to the W. African (254,364) and the Cameroon reference panels (435,206) Table 4.1. A two-fold increase was observed in the Cameroon reference panel compared to the W. African reference panel, despite the reduced number of samples in Cameroon (N = 9) compared to W. Africa (N = 21) (Table 4.1). The sites in the reference panels were biallelic, any sites with missing data were removed, and all sites were phased prior to imputation. After filtering the target population for minimum depth ≥5, MAF count ≥ 3 and <50% missingness with *vcftools*, 195,085 and 174,877 variant sites remained in the W. Africa target 1 (N = 66) and the W. Africa target 2 (N = 45) (a subset of the whole genome sequences with uneven coverage obtained from Grant Lab) populations respectively, while 37,107 and 39,009 variant sites remained in the Cameroon target 1 (N = 86) and Cameroon target 2 (N = 95) populations. These were the number of variant sites present in the populations before carrying out *beagle* filtering steps with *conform-gt* (Table 4.1).

I observed that the count of variable sites that met the *beagle* filtering with *conform-gt* criteria (that is, the step which ensures that the variant sites in the target are also in the reference) in the pre-imputation target populations was greater in the W. African target populations compared to the Cameroon target populations (Table 4.1): there were ten-fold more sites meeting *beagle* filtering with *conform-gt* criteria amongst the variant sites in the

W. African target population compared to the Cameroon target population, probably because the W. African reference pop is larger. The number of variable sites after running *beagle* filtering with *conform-gt* step in the W. African target 1 (N=66) and W. African target 2 (N=45) populations were 136,784 (68.66% of the total) and 105,198 (60.16%), respectively (Table 4.1). Fewer than 4% failed due to inconsistent or inconclusive chromosome strand evidence in both W. African target populations, and 27.40% and 36.40% of the variant sites in W. African target 1 (N=66) and W. African target 2 (N=45) populations were removed because they were not present in the reference variant calls.

In contrast, in the Cameroon target populations, less than 30% of the total variant sites remained after running the *beagle* filtering with *conform-gt* step (that is, the count of variant sites that passed) in both target populations (Table 4.1). Following *beagle* filtering with *conform-gt*, greater than 70% of the data were removed because they were not present in the reference variant calls, while <2% failed because of inconsistent or inconclusive chromosome strand evidence.

Table 4.1. Counts of variant sites in the reference panels, the counts, and the percentage (in bracket) of variant sites of the target populations before and after running *beagle* filtering with *conform-gt* steps, and after imputation.

Reference panel	Variant sites in the reference panel	Target population	Target population variant sites that passed *vcftools* filtering	Target population variant sites that are present in the reference and the target (%)	Target population variant sites after imputation
global panel (N = 27)	569363	West African (N = 66)	201365	136784 (68.66)	569363
		Cameroon (N = 95)	39009	10333 (26.49)	569363
West African panel (N = 21)	254364	West African (N = 45)	174877	105198 (60.16)	254364
Cameroon panel (N = 9)	435206	Cameroon (N = 86)	37107	9885 (26.64)	435206

NB: All variant sites are biallelic.

The variant sites in the global panel (N = 27) (whole genome sequences obtained from Choi et al. (2016); average depth of coverage = >20) was used to impute the W. African (N = 66) (whole genome sequences with uneven coverage obtained from Grant Lab) and Cameroon (N = 95) target populations (reduced representation sequences obtained from Grant Lab). The variant sites in the W. African panel (N = 21) have average depth of coverage of >20. While the the variant sites in the Cameroon panel (N = 9) have average depth of coverage >5.

The target population variant sites that are present in the reference and the target (%) column contains the number of “confirmed” variable sites in the target population prior to imputation (those are in the variant sites that passed the *beagle* (*conform-gt*) filtering step).

4.3.2. Accuracy and yield of imputation from a reference panel with diverse population ('global') compared to reference panels derived from the same population as the target samples ('W. African' and 'Cameroon').

The yield and accuracy of imputation on low-depth whole genome sequence data, and on reduced representation genome data, were assessed using reference panels from (1) a more diverse population and (2) a subset of same population, as the target populations. I used filtered whole genome sequences from a population sample of 27 individuals containing 569,363 variable sites as the global reference for imputing the W. African (N = 66) and the Cameroon (N = 95) target populations, separately (Figure 4.1). Likewise, I used a subset (N = 21) of the W. African worm population as a reference population for imputation of missing data from a W. African target population. The W. African reference population was composed of W. African worms that had been sequenced at higher depth (>20; 254,364 variable sites), while the target population was composed of W. African worms sequenced at a low depth (<5) (Figure 4.1). Similarly, I used a subset (N = 9) of the Cameron worm population for which there were whole genome data (sequenced at depth >5; 435,206 variable sites) as a reference to impute genotypes for a sample (N = 86) from the same Cameroon worm population that had been sequenced using a reduced representation method (Table 4.1; Figure 4.1).

4.3.2.1. Comparison of imputation yields across all populations.

Allele frequency spectra or distribution (Figure 4.2) of the reference panels were compared with the allele frequency spectra of the pre- and post-imputation data. Subjectively, the figure shows that the reference and post-imputed allele frequency spectra were not different, but a Wilcoxon signed rank test showed that there were statistically significant differences between the reference and post imputed allele spectra for all comparisons (Table 4.2).

The genotype imputation yielded the same counts of variable sites in the post-imputation data as that of the reference panel used for the imputation as shown in Table 4.1 such that the more diverse the samples in the reference panel, the greater the imputation yield. For example, the global reference panel yielded more variable sites in the target populations after imputation (global/W. Africa = 569,363; global/Cameroon = 569,363) compared to the target populations imputed from a reference panel of the same population (W. Africa/W. Africa = 254,364; Cameroon/Cameroon = 435,206) (Table 4.1).

Imputation changed the allele frequency in the target population and improved the ability to detect likely rare variants (variants with MAF <0.01). Comparing the pre-imputation frequency spectrum with the post-imputation frequency spectrum for both imputation trials (that is, the separate imputations carried out with different reference panels), an interesting difference is the degree to which there was an increase in numbers of rare variants in the post-imputation target population relative to the pre-imputation target population. Both the global/W. Africa and W. Africa/W. Africa combinations (Figure 4.2 A&B) had an approximately 20% increase in the rare variants in the target population following imputation (relative to the frequency spectrum of the pre-imputation target population) using either reference panel or the common variants were also increased (MAF >0.05; global/W. Africa = 75.89 % increase; W. Africa/W. Africa = 74.17% increase) (Figure 4.2 A&B). Comparably, global/Cameroon and Cameroon/Cameroon imputation combinations yielded approximately 99.20% and 91.72% increase in the rare variants in the target population following imputation with the two reference panels, and an approximate 98% increase in the common variants for both (Figure 4.2 C&D).

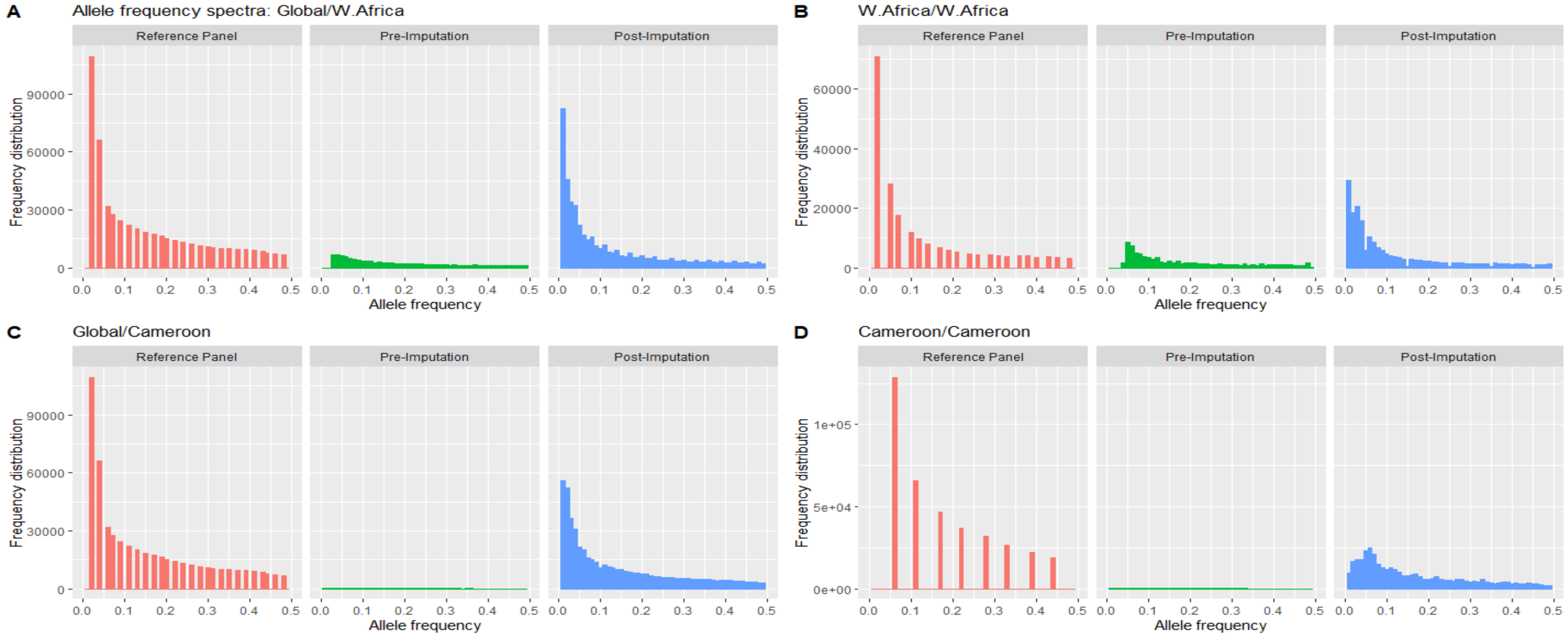


Figure 4.2: Histogram plot showing the allele frequency spectra for the reference panel (red) vs post-imputation (blue) for all imputation combinations.

The figure shows the allele frequency distribution for variant sites imputed by the A) global reference panel/W. African target population; B) W. African reference panel/West African target population; C) global reference panel/Cameroon target population; and D) Cameroon reference panel/Cameroon target population. See also Figure 4.1. The x-axis is the allele frequency values for the variant sites while the y-axis shows the frequency distribution of variant site at each value. Each bar in the histogram plot has a binwidth of 0.01.

Table 4.2. Table of Wilcoxon signed rank test statistics for all imputation set comparisons.

Reference/Post-Imputation data	Wilcoxon signed rank test value	p-value
global/W. Africa	1.17e^{+11}	< 2.2e^{-16}****
global/Cameroon	1.03e^{+11}	< 2.2e^{-16}****
W. Africa/W. Africa	2.35e^{+10}	< 2.2e^{-16}****
Cameroon/Cameroon	5.82e^{+10}	< 2.2e^{-16}****

A table showing the results from Wilcoxon signed rank test on the four imputation combinations (global reference panel/W. Africa target population, global reference panel/Cameroon target population, W. Africa reference panel/W. Africa target population, Cameroon reference panel/Cameroon target population). It shows that there was significant evidence of a difference between the reference and imputed allele spectra for all comparisons. it is also of interest to note that the rank test value (although not p-value associated with it) is lower for the W. Africa/W. Africa and Cameroon/Cameroon.

4.3.2.2. Comparison of imputation accuracy across all populations.

Beagle v4.1 provides two position-level imputation metrics: allelic R-squared (AR2) and dosage R-Squared (DR2). Dosage R-squared (DR2) is defined as the estimated squared correlation between the estimated allele dose and the true allele dose. AR2 is defined as the squared correlation between the allele dosage of the most likely imputed genotype in the target population and the true allele dosage in the reference (allele dosage in this context is the number of minor alleles). Both AR2 and DR2 have been indicated to be good measures of imputation accuracy and can be used to filter imputed variants prior to downstream analysis. The two measures of imputation accuracy (AR2 and DR2) reported by *beagl*e were compared and Figure 4.3 shows that they are strongly correlated across all combinations of reference-target imputation combinations: global/W. Africa and W. Africa/W. African (r^2 = 0.99, p < $2.2e^{-16}$) compared to the global/Cameroon and Cameroon/Cameroon (r^2 = 0.97, p < $2.2e^{-16}$). The heat map colours on Figure 4.3 indicate allele frequency (AF) of the imputed data. The red colours (low AF) tend to be further away from the line (where AR2=DR2) and lighter colours (mid-range AF) are closer to the line, which indicates that the correlation between AR2 and DR2 is dependent on AF. (Figure 4.3). Better correlation was observed at moderate frequency (more green tones closer to the line as AR2 increases) for all four combinations, especially for Cameroon. This is expected based on (a) the confidence that imputation should be better for moderate allele frequencies and (b) the reference panel and target populations are larger for W. Africa in particular, which also means higher imputation confidence even at lower AF (more red closer to the line and at higher AR2 in Figure 4.3 A&B).

In this study, the measure of how confident one can be that the imputed allele at any given variant site is likely to be true (that is, a measure of the confidence in the identity of the imputed allele) was assessed by AR2. AR2 ranges between 0 and 1. A high correlation (values closer to 1) between allele dosage in the target population and dosage in the reference means that the reference and target populations are similar. Low correlation (values closer to 0) implies that the target and reference populations are different. The AR2 metric was chosen for pruning the variants before performing GWAS. There are different approaches to exploring the relationships between AR2 and imputation outcomes and I assessed that using the following four criteria.

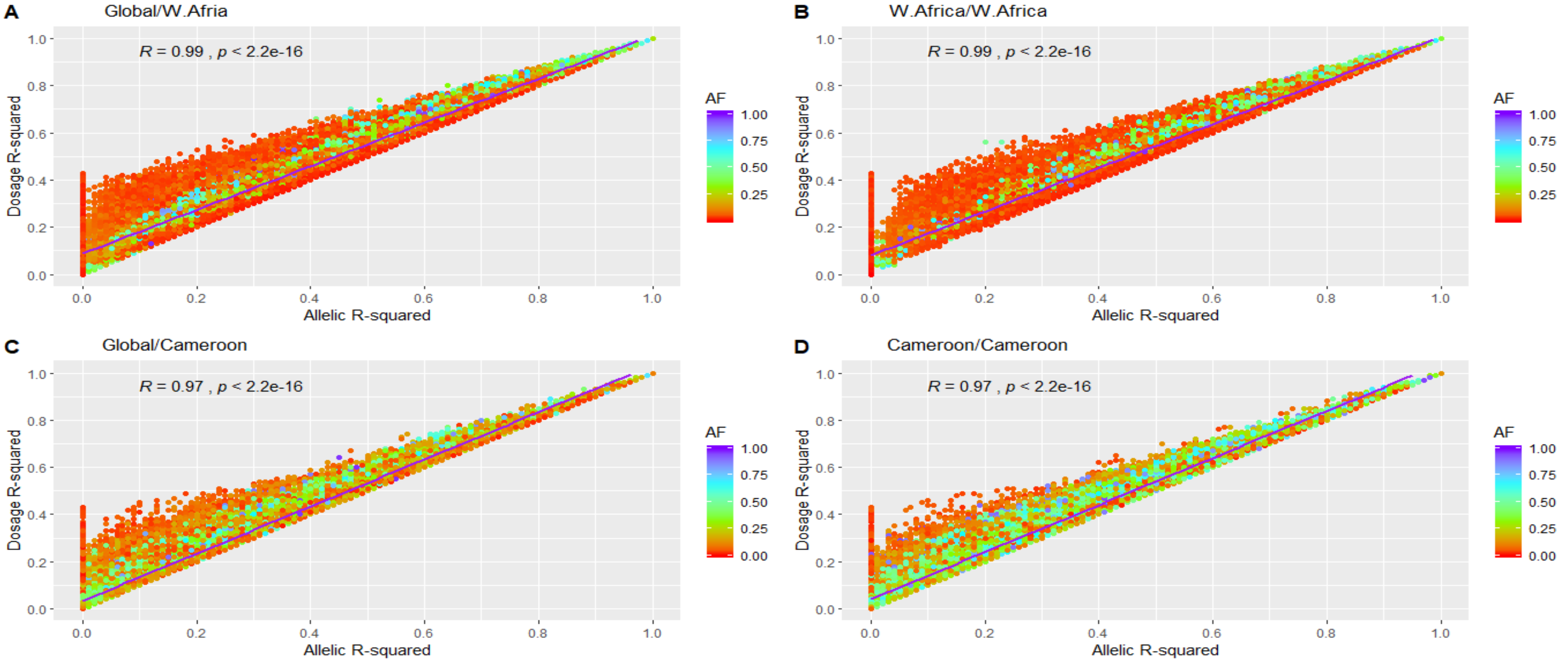


Figure 4.3. The relationship between two imputation accuracy measures, allelic R-squared and dosage R2 ranked by allele frequency (AF) values.

A) global reference panel/W. African target population. B) W. African reference panel/W. African target population. C) global reference panel/Cameroon target population, and D) Cameroon reference panel/Cameroon target population. The x-axis is the Allelic R-squared while the Y-axis is the Dosage R-squared. Each coloured point corresponds the distribution of the variants according to their allele frequency values. The figure shows that these two measures of imputation accuracy are strongly correlated. The correlation between them is best for high frequency imputed alleles (AF > 0.1). For all four combinations, the correlation is better at moderate frequency (more green tones closer to the line as AR2 increases), especially for Cameroon.

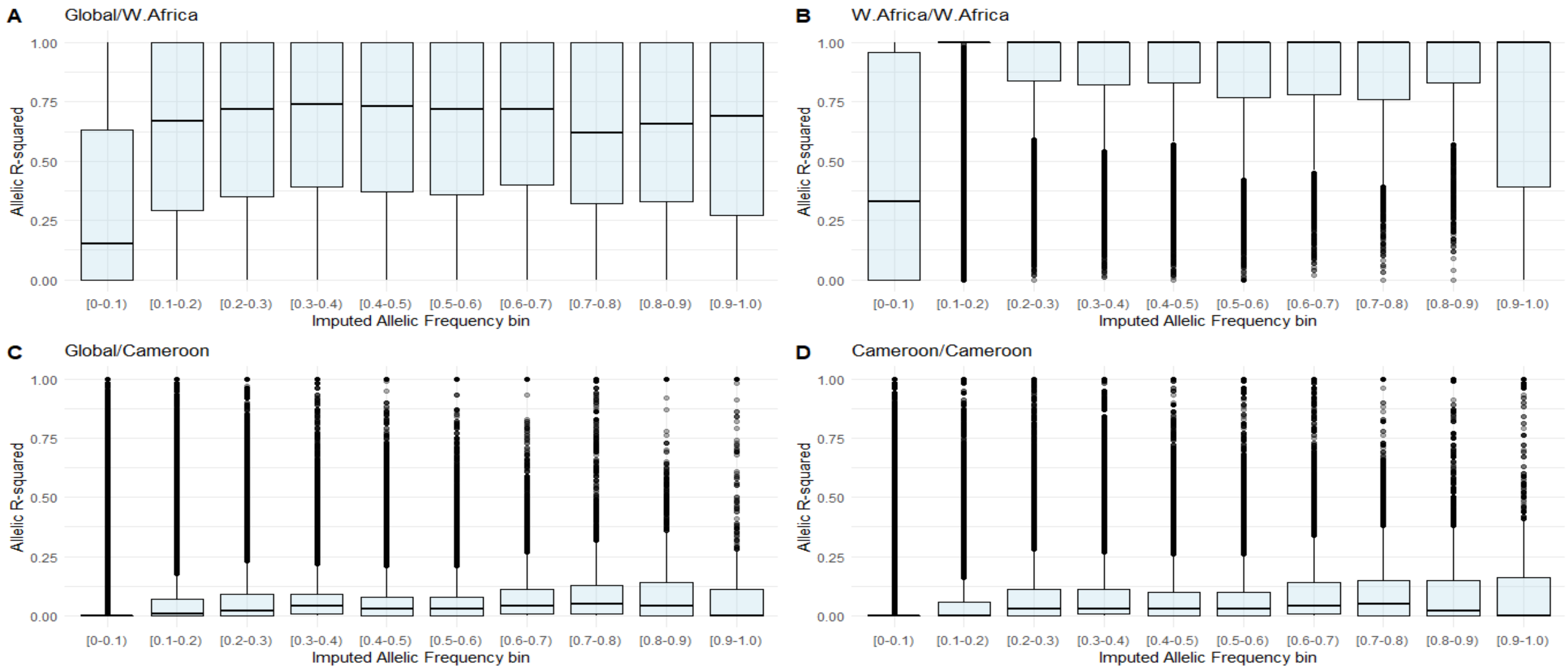


Figure 4.4. The distribution of allelic R-squared (AR2) across 10 allele frequency bins for imputed variants.

The X-axis shows 10% frequency bins: [0,0.1), [0.1,0.2) …, [0.9,0.1). The Y-axis shows the distribution of AR2 in each AF bin for the four reference/target population imputation combinations: (A) global/W. African; (B) W. African/West African; C) global/Cameroon; and D) Cameroon/Cameroon. The boxplot shows statistical bounds of 1st and 3rd quartiles at the lower and the upper horizontal lines, respectively. The horizontal line within the box is the median value, and the lines extending above and below the boxes (whiskers) are the outlier data points.

Assessment criterion one: the distribution of AR2 across various allele frequency bins. The first method used to assess the imputation accuracies of the combinations of reference panel and target population was to compare the distribution of AR2 across various allele frequency bins. Figure 4.4 shows the distribution of AR2 across 10 alternate allele frequency bins for the imputed variants. Variation in the value of AR2 was observed for variants in the global reference/W. African target population compared to the W. African reference/W. African target population, particularly especially, with the common variants (AF > 0.1) (Figure 4.4 A&B). For global/W. African and W. African/W. African imputation combination, AR2 is lower at AF<0.1, and uniform above 0.1. The difference in the range of AR2 for W. African/W. African imputation combination is smaller and the AR2 values higher (smaller boxes higher up the y-axis). The Cameroon imputed populations have quite different relationships with allele frequencies. Cameroon population has uniformly low AR2 for all allele frequencies irrespective of the allele frequency bin.

Since AR2 is a measure of imputation accuracy, Figure 4.4 implies that the imputation of W. African populations has AR2 > 0.7 provided AF>0.1, but that AR2 is <0.1 for all values of allele frequency in the Cameroon populations. That is, the accuracy of the Cameroon population imputation is poor relative to the W. African population.

Assessment criterion two: Correlation of the imputed allele frequencies with the reference allele frequencies. The second method I used to compare imputation accuracies of the combinations of reference panel and target population was by examining the correlation between the observed minor allele frequency from the reference panel (that is, the un-imputed minor-allele frequency) and the minor allele frequencies estimated from the imputed data as shown in figure 4.5. There were strong correlations between the minor allele frequencies estimated from reference and minor allele frequencies in imputed target populations across all imputation combinations using Pearson's r: global/W. Africa ($r^2 = 0.81$, p-value < $2.2e^{-16}$); W. Africa/W. Africa ($r^2 = 0.94$, p-value < $2.2e^{-16}$); global/Cameroon ($r^2 = 0.94$, p-value < $2.2e^{-16}$); Cameroon/Cameroon: ($r^2 = 0.90$, p-value < $2.2e^{-16}$) (Figure 4.5).

The statistical testing of the allele frequency spectra supports a conclusion that there are statistically significant differences between the reference and target population frequency spectrum (Table 4.2), which is at odds with the superficial interpretation of Figure 4.5. Figure 4.5 supports a conclusion that the correlation between reference and imputed frequencies is strongly dependent on allele frequency, and that dependency is very different between the W.

African imputations and the Cameroon imputations. Thus, the Cameroon imputations appear to have strong correlation between reference and imputed, but the fact that this is true only at low allele frequency is of concern. High correlation between the minor allele frequencies estimated from reference and minor allele frequencies in imputed target populations implies that although they are statistically different as suggested in Table 4.2, the difference is not biologically significant. This may be because all the spectra are strongly over dispersed and are skewed strongly towards the low frequency alleles.

Imputation accuracy is lower in the global/Cameroon and Cameroon/Cameroon (Figure 4.5 C & D): AR2 had low values across the entire MAF (lots of red dots across the plot). The reference allele frequency did not influence the imputation quality measures in the global/Cameroon and Cameroon/Cameroon population because there was no correlation between the AR2 measures and the allele frequencies (Pearson correlation coefficient (r^2) = 0.05 at p-value < $2.2e^{-16}$ vs 0.07 at p-value < $2.2e^{-16}$). The p-values showed that even though these correlations were relatively low they had strong statistical support. There was a strong effect of allele frequency on AR2 for W. Africa (this is also evident from Figure 4.4,): the allele frequencies were slightly correlated with AR2 in the global/W. African and W. African/W. African populations (Pearson correlation coefficient (r^2) = 0.37 at p-value < $2.2e^{-16}$ vs 0.44 at p-value < $2.2e^{-16}$). The p-values indicated that even though the correlations were relatively low they have strong statistical support, which might be from the large number of observations considered. While there is no correlation for Cameroon populations.

Figures 4.4 and 4.5 make it clear that (a) the confidence in the Cameroon imputations is low and (b) that imputation quality is much better for W. Africa but is also allele frequency dependent for W. Africa (low confidence at low allele frequency). The overall result is that the allele frequency distributions are generally similar, but that the Cameroon imputations are very poor at all allele frequencies (whereas the W. Africa imputations are of higher quality for all but a proportion of low frequency alleles). Furthermore, it was at low frequency that the difference between Global and W. Africa reference panels for W. Africa imputation got clearer (the W. African reference did better at low frequency).

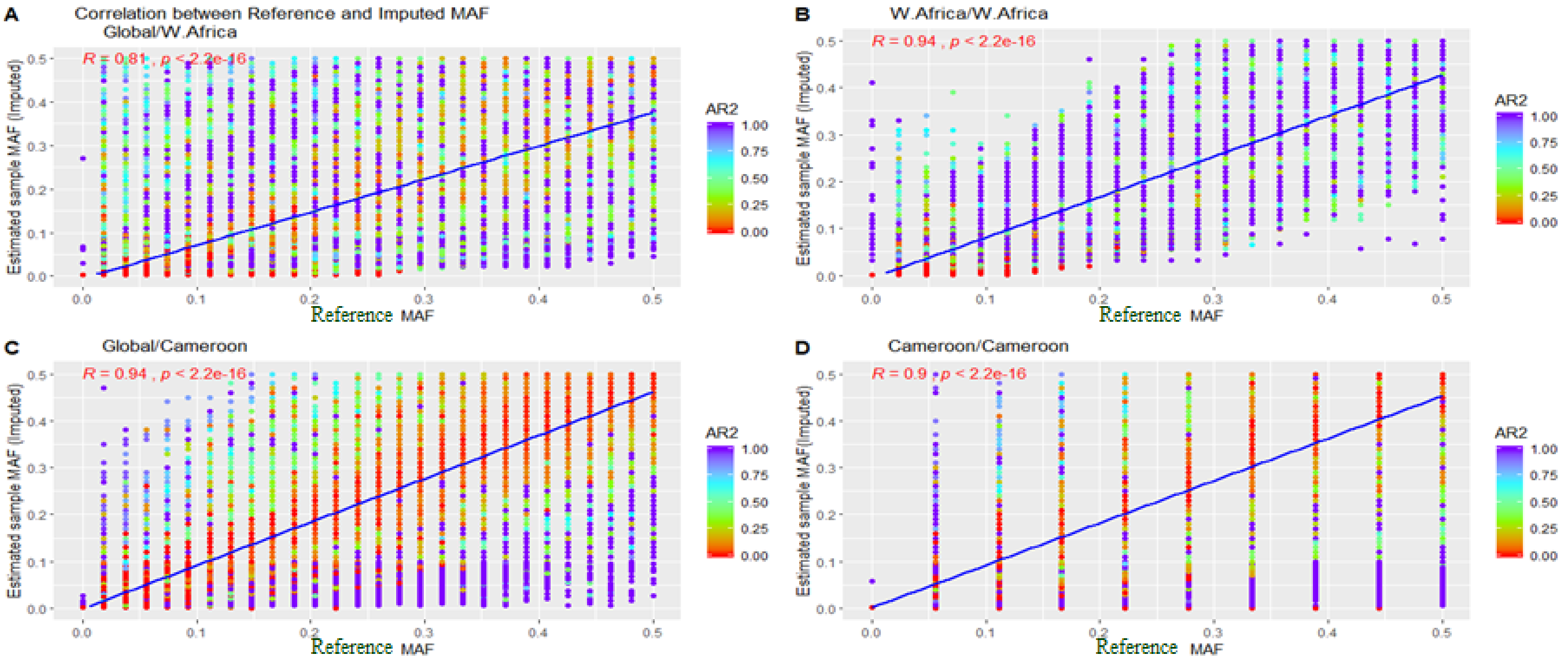


Figure 4.5: Correlation between the reference and imputed sample minor-allele frequency (MAF) ranked by AR2 values.

The figure shows the correlation between the Reference MAF (on X-axis) and the imputed MAF (on Y-axis) in (A). global reference/W. African target population; (B). W. African reference/W. African target population; (C). global reference/Cameroon target population; and (D). Cameroon reference/Cameroon target population. The colour coding shows the AR2 of each data point. It is interesting that A & B are very different to C & D. For A & B, the data points closest to the line (that is, the best correlation) have high AR2, whereas the opposite is true for C & D.

Assessment criterion three: application of an AR2 filter: An estimate of AR2 > 0.80 is usually recommended as a filtering score to apply to the imputed data before performing GWAS (H. Daetwyler, pers. comm.). Following application of this threshold, 163,845 (28.78%) and 119,914 (47.14%) of imputed alleles passed the AR2 filtering in global/W. African and W. African/W. African imputation combinations, respectively. W. African reference is more appropriate for W. African imputation. In contrast, only 13,074 (2.30%) and 12,245 (2.81%) passed the AR2 filtering in the global/Cameroon and Cameroon/Cameroon imputation combinations, respectively. Compared to the number of sites that passed the beagle filtering stage prior to imputation from Table 4.1, the following number of variable sites were added to the reference/target combinations: global/W. Africa = +27,061, W. Africa/W.Africa = +14,713; global/Cameroon = +2741, Cameroon/Cameroon +2325.

Assessment criterion four: comparison of high/low AR2 allele distribution. The distribution of allele frequencies for the imputed variants that have AR2 > 0.8 was compared with the allele frequencies for the imputed variants with AR2 < 0.8 to see if there are differences in allele frequencies above or below the threshold of AR2 > 0.8. Figure 4.6 shows the allele frequency spectrum of the imputed variants and their categories by AR2 values. The variants having high imputation accuracy, that is, AR2 above 0.8, and those with AR2 below 0.8 show frequency spectra that were distributed similarly (Figure 4.6).

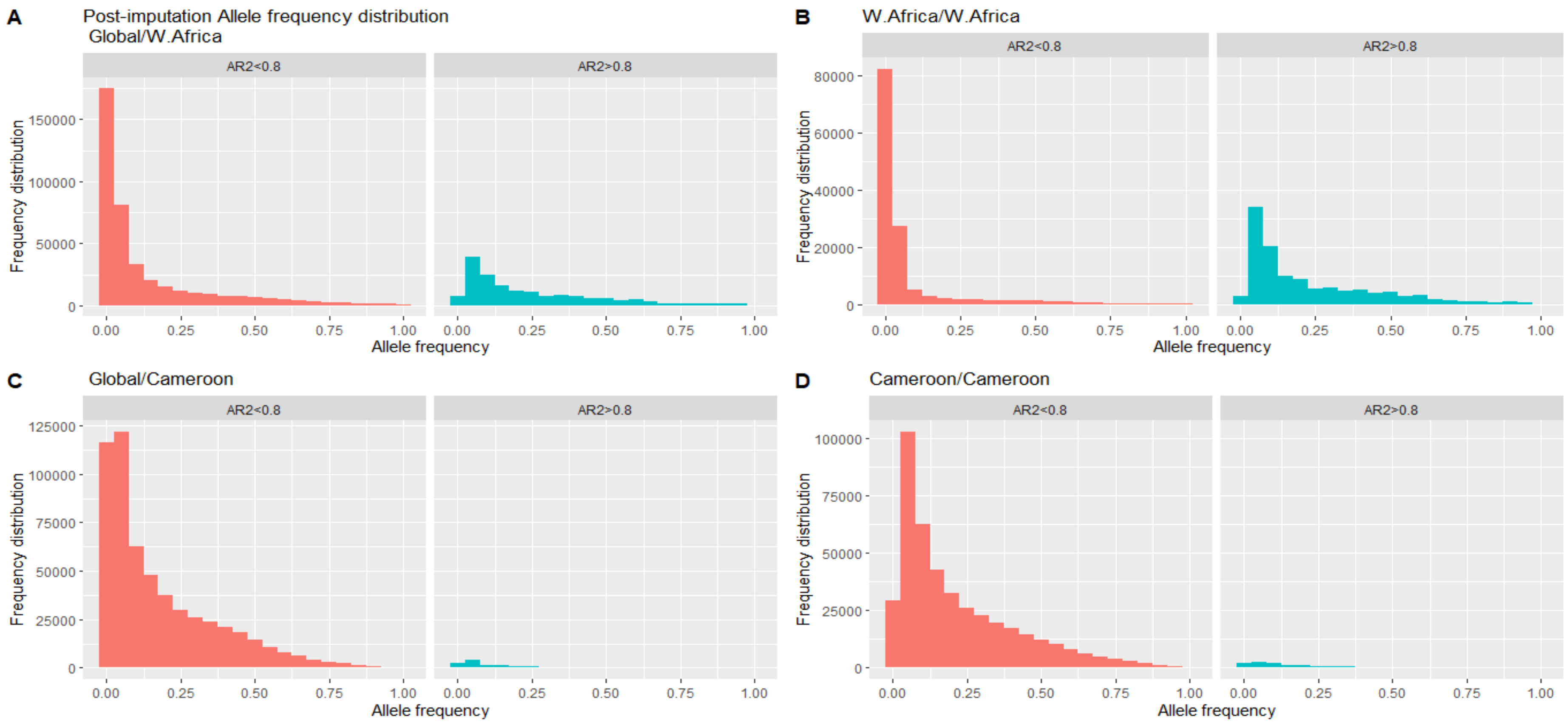


Figure 4.6: Allele frequency spectrum for each of the four combinations of reference/target populations.

The figure shows the distribution of allele frequencies for the imputed variants that have 0.8 ≤ AR2 ≥ 0.8 (red and blue histogram plot respectively) for (A) global reference panel/W. African target population; (B) W. African reference panel/West African target population; (C) Global reference panel/Cameroon target population; and (D) Cameroon reference panel/Cameroon target population. Each histogram plot has a binwidth of 0.05 (*x*-axis; frequency). This figure shows that, as expected, filtering by AR2 removes mainly the lowest frequency alleles from the W. African data but removes almost everything from the Cameroon data. It reinforces the point that there is an effect of allele frequency on imputation quality.

4.3.3. Association analysis.

Genome-wide association tests were carried out using pre- and post-imputation genotypes for the W. African and Cameroon target populations to determine whether imputation improves the power of the association of the variants with ivermectin response.

After carrying out QC steps to filter out imputed variant sites of AR2 ≤ 0.8 and individuals based on phenotypes (non-phenotyped worms from the W. African population and day 0 worms from the Cameroon population were excluded); 13,073 variant sites and 65 worms passed in the global reference panel/Cameroon target population imputation combinations; 12,244 variant sites and 56 worms passed in the Cameroon reference panel/Cameroon target population imputation combinations. In contrast, 163,844 variant sites and 59 worms passed in the global reference panel/W. African target population imputation combinations; and 119,914 variant sites and 38 worms passed in the W. African reference panel/W. African target population imputation combinations.

As an essential part of the QC steps, the presence of population stratification was tested using a multidimensional scaling (MDS) approach on the imputed data for the Cameroon and the W. African populations. The aim was to reveal groups of individuals that are genetically more like each other than expected. Figure 4.7 shows no structure between worm phenotypes from the same target populations (that is, worms from individual target population are genetically similar or identical) but the worms from different populations (W. Africa and Cameroon) are genetically different (Figure 4.7). This is consistent with previous studies (Choi et al. 2016; Doyle et al. 2017; Crawford et al. 2019).

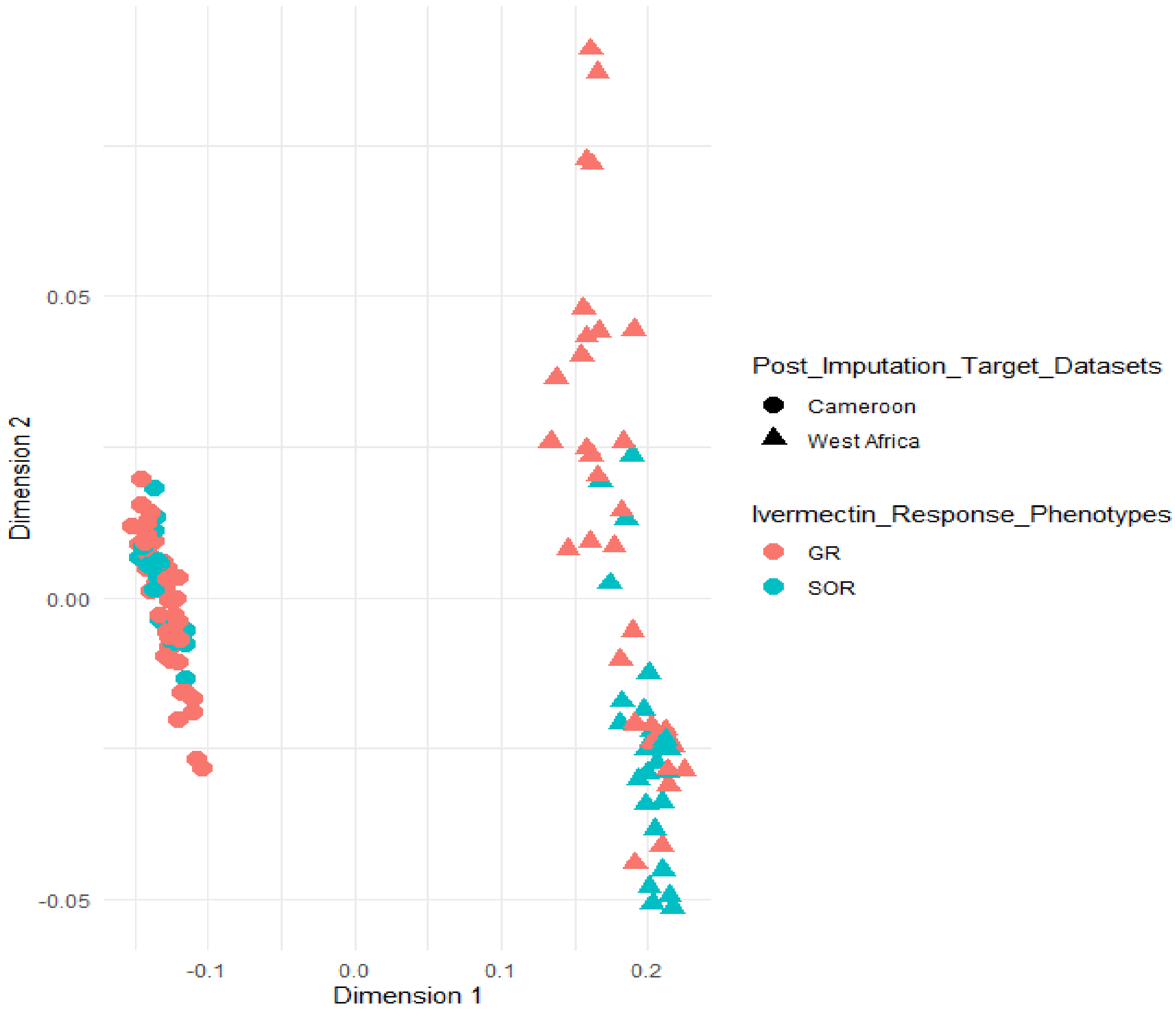
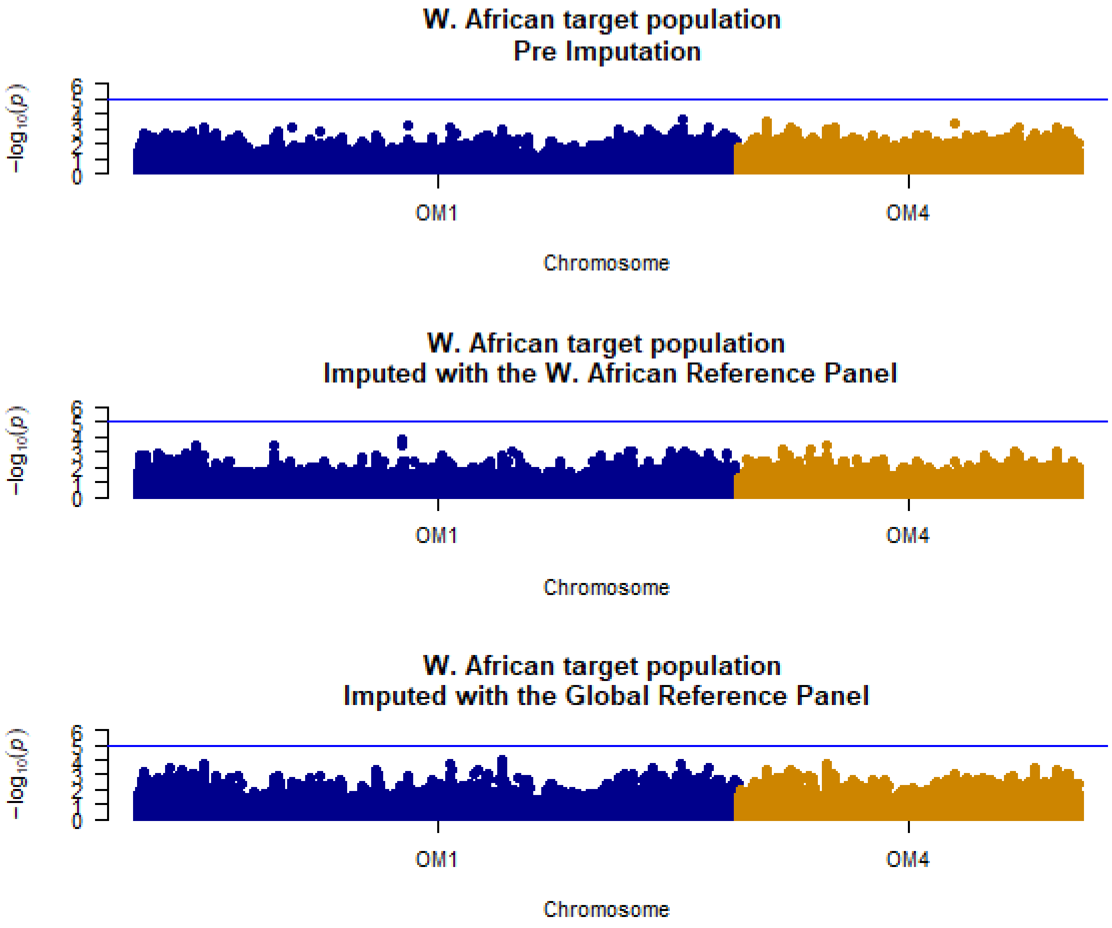


Figure 4.7. Non-metric multidimensional scaling plot, based on genetic data, to identify population stratification (or relationships) between worms from Cameroon/West Africa with good (GR) or poor (SOR) ivermectin response phenotypes used in the association study.

The circles and triangles are the samples from Cameroon and W. Africa respectively, while the red and the blue colour corresponds to their respective phenotypes (GR and SOR).

Figure 4.8 shows the GWAS result for the W. African populations imputed with the W. African and the global reference panel. When compared with the pre-imputation GWAS, the strength of genetic associations with the phenotype (SOR) increased following imputation from the global reference panel/W. African target population combination (Figure 4.9, lower panel). The figure shows that no variant site was significant genome-wide after correcting for multiple testing, that is, there was no association with drug phenotype that met a Bonferroni correction threshold of p-value < 0.05 in the global/W. African and W. Africa/W. African imputation combinations. The reason for no significant association may be, perhaps, Bonferroni was too stringent. Similar values were obtained with Benjamin-Hochberg corrections.

Figure 4.9 shows the GWAS result for the Cameroon populations imputed with the Cameroon and the global reference panel. Approximately 500 variant sites showed a statistically significant association genome-wide with ivermectin response in the pre-imputed data (the reduced representation data) after correcting for multiple testing (Bonferroni correction p-value < 0.05). This contains lots of noise. Following imputation in the Cameroon and the global reference panels, 10 SNP loci were statistically significant after correcting for multiple testing (Bonferroni correction p-value < 0.05).



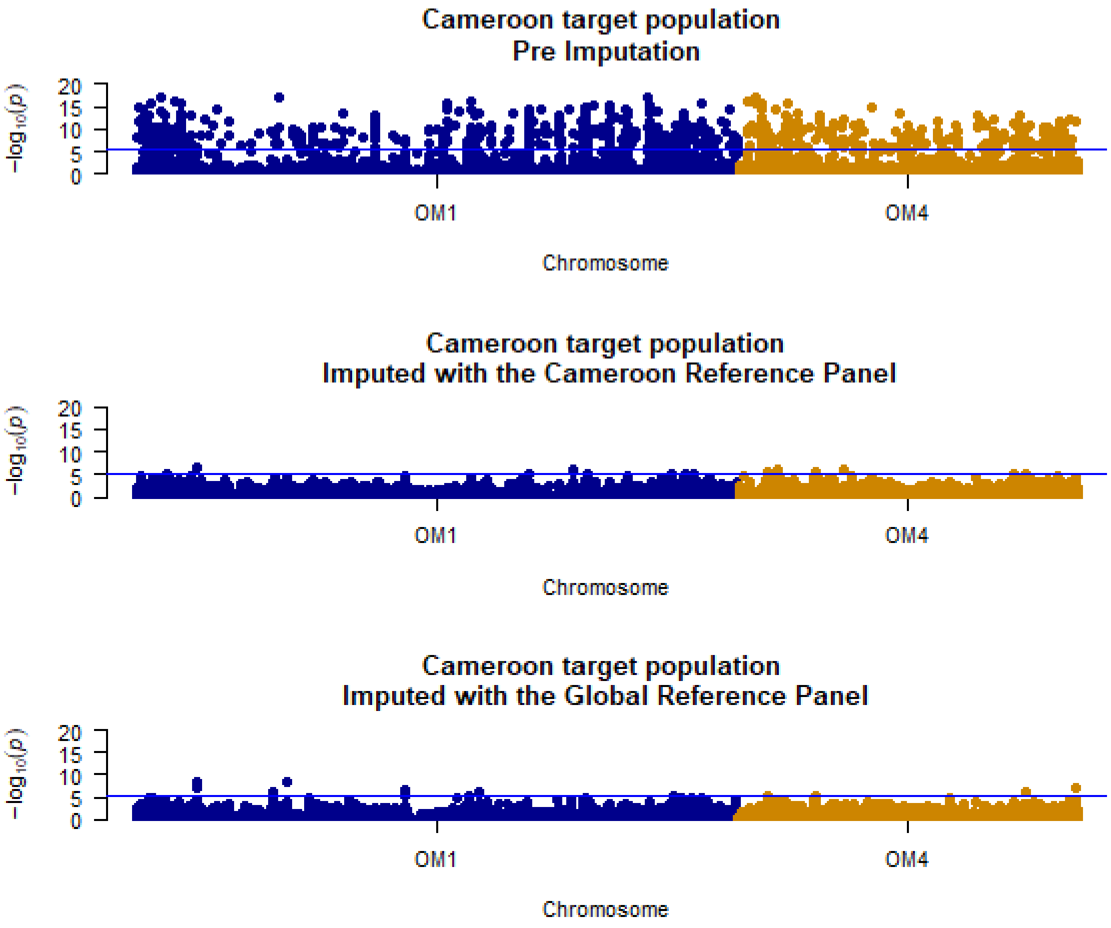
W. African target population
Pre Imputation

W. African target population
Imputed with the W. African Reference Panel

W. African target population
Imputed with the Global Reference Panel

Figure 4.8. Manhattan Plot showing GWAS result for the W. African target populations pre-imputation (top), and post-imputation with the W. African reference panel (middle) and Global reference panel (bottom).

The X-axis represents the position of a variable site on the autosomal chromosomes OM1 and OM4. The Y-axis is the relative -log10 p-values of each variant site. None of the associations were strongly associated with ivermectin response after Bonferroni correction. The blue horizontal line is the Bonferroni significant threshold of P < 0.05.



Cameroon target population
Pre Imputation

Cameroon target population
Imputed with the Cameroon Reference Panel

Cameroon target population
Imputed with the Global Reference Panel

Figure 4.9. Manhattan plot showing GWAS results for the Cameroon target population pre-imputation (top), post-imputation with the Cameroon Reference (middle), and Global Reference (bottom) panels.

The X-axis is the position of each SNP along the autosomal chromosomes OM1 and OM4. The Y-axis is the relative -log10 p-values of each variant site (the dots). The blue horizontal line is the Bonferroni significant threshold of $P < 0.05$. The points above the blue lines show very strong association with ivermectin response because they were significant after Bonferroni correction.

Discussion

For the first time in filarial nematodes, I have described the concept and a practical strategy for imputing genotypes from a reference panel composed of either a diverse population or one from the same population as the target population. Overall, my results showed that the imputation from a reference panel derived from the same population as the target data (precisely, the W. African reference panel) is superior to the imputation from a genetically and geographically diverse population (the global reference panel) because the W. African /W. African imputation combination gave better accuracy results (AR2) than imputation from the global reference panel.

The major caveat of this study is the small sample size in both reference panels (global = 27, W. Africa = 21 and Cameroon = 9) and target populations (W. Africa = 66 and 45; Cameroon = 96 and 85) relative to those in the organisms for which these methods have been developed. Previous studies reported imputation with and on large samples sizes (for example, 100 individuals and above) (Browning and Browning, 2009, Browning and Browning, 2016, Das et al., 2016, Howie et al., 2012, Li et al., 2009, Marchini and Howie, 2010) among others. Isik et al. suggested that a larger number of sequenced individuals from a species is required to use as reference haplotype panels (Isik et al., 2017). The aim of the work reported in this chapter was to obtain the first elements towards proof-of-concept via an exploration of genomic imputation as an option for filarial nematodes using the available data. Those first elements towards proof-of-concept have now been demonstrated but the limitations of sequencing depth exist, and also the fact that it is a human genetics tool that does not necessarily deal properly with helminth population dynamics, hence, the challenge of developing the genomic resources to take advantage of this technology.

When a reference panel for a population is not available, one can choose a subset of the target sample with a more complete genotype data as the reference panel (for example, as done here with the W. African reference panels) and use that more densely genotyped subset to impute the variants for the remainder of the sample (Anderson et al., 2008). This has the additional advantage that the reference panel is perfectly matched to the target sample (Browning and Browning, 2009). My study is consistent with this observation in that the W. African reference panels did well in terms of accuracy compared to a global reference panel. For example, the W. African reference is better by AR2 distribution for all allele frequencies (Figures 4.4 to 4.7), and the correlation of pre- and post-imputation allele frequencies was stronger for the W. Africa reference compared with the global reference (Figure 4.5). Also, there were twice as many sites that passed the AR2>0.8 filter in the W. Africa/W. Africa imputed data than in the global/W. Africa data and the mean imputation accuracy of variants with MAF >0.05 was higher in the W. Africa/W. Africa population compared to the global/W. Africa data (AR2 0.737 *vs* AR2 0.550 respectively) (Appendix Table 4.1). This is in contrast to the Cameroon imputations where the AR2 performance was poor regardless of which reference panel was employed. This suggests that the limitation with the Cameroon imputations lies with the nature of the target population genotypes, which for Cameroon was composed of reduced representation sequence data (NuGen Allegro). The reduced representation sequencing yielded approximately 10^4 SNP loci, with an average $10^3 – 10^4$ bp between loci. The LD threshold for imputation is a r^2 >0.33 (chapter three), which permits accurate imputation over distances of <1.5 kb in the *O. volvulus* genome (chapter three). The imputation outcome, therefore, is that the reduced representation genotype data are too sparse to permit genome wide imputation and the yield of “useable” imputed genotypes is therefore small.

4.4.1. Impact of composition, diversity, and size of reference panels and target populations on imputation accuracy.

One important deduction from this study is that the choice of the reference panel has a substantial impact on measures of imputation accuracy, particularly when imputing low-frequency genetic variants (that is, variants with MAF <0.05). I demonstrated with the W. African reference/W. African target population imputation combination that the use of the same reference panel as the target population produces substantial gains in imputation accuracy as seen in appendix Table 4.1 and Figure 4.4 – 4.6. The relationship between mean AR2 and MAF (appendix Table 4.1) and the distribution of allele frequencies for the imputed variants that have AR2 > 0.8 (Figure 4.6) indicates that genetic variants with a frequency as low as 0.01 could be imputed when using a diverse reference panel, but they are not useful for GWAS because their AR2 values are too low.

The origin, size and the diversity of the reference panel influenced the measures of AR2 according to the study of Browning and Browning (2009). They reported an increase in imputation accuracy with increase in reference panel; that is, measures of AR2 increased as more individuals were added into the reference panel. The contrary was observed in my study, lower imputation accuracy with the global reference panel, (global = 27, W.

Africa = 21 and Cameroon = 9) (Figure 4.1) even though there was increase in yield of variants, especially low/rare-frequency variants (MAF <0.05). Two important factors could have impacted this: the first and the most important factor is population structure – the global reference panel consists of worms from multiple populations, representing a larger slice of genetic diversity across the *O. volvulus* species - which implies that many of the low/rare alleles in the global reference will likely be population specific (or their frequency in different populations may vary: this was the pattern observed when comparing alleles and allele frequencies between W. African and Cameroon). This also means that there would be an increase in the number of rare/low frequency alleles in the imputed data that are not present in the population, hence the low AR2 values. A second but also important factor is that the ability to detect rare/low frequency alleles in a population is a function of sequencing depth. This means that rare/low frequency alleles will be underrepresented in the target population because of the problems with sequencing depth in that population. The outcome of imputation is that those "missing" rare/low frequency alleles will be overrepresented in the imputed data because they are underrepresented in the experimentally derived target data. This also means higher "fake" homozygosity that will certainly alter DR2 and AR2 values.

Variation in the composition and quality of the datasets used in this study could also influence the variation in the number of variants observed and their outcomes in terms of imputation accuracy. The global reference panel is composed of high-quality sequences with high sequencing depth that are derived from single worms covering the entire global range of *O. volvulus*. The higher number of variants observed in the global reference compared to the W. African (or Cameroon) references is in part because they were genotyped at greater depth (and will thus detect more rare and low frequency alleles) (Choi et al., 2016) and in part because the geographic diversity of the samples present means more variation across many parasite populations is captured. The W. African reference panel is also composed of good quality sequences with high depth of coverage (>20) but is primarily drawn from a single (Ghanaian) population that is likely to contain only a subset of the global variation (Choi et al., 2016). The Cameroon reference panel are composed of whole genome sequence but only at a minimum average depth of coverage of 5. The high number of variants (approximately two-fold increase) observed in the Cameroon reference panel compared to the W. African reference panel is likely because the filtering parameters for the Cameroon data had a lower minimum depth requirement for variant calling or because of higher genetic diversity in the Cameroon population compared to the W. African population. Higher genetic diversity of worms from Cameroon than those from Ghana (which takes the largest part of the W. African target population) has been established previously by Doyle et al (2017).

Although the W. African target population was also composed of generally good quality whole genome sequences, they were intermixed with regions of low quality in terms of depth and coverage. This impacted the difference in the number of variant sites observed between them after *vcftools* filtering (Table 4.1). Depth is the number of reads at a given genomic position, coverage is the proportion of the genome that is represented in the sequence. These two parameters are often correlated (more depth usually means better coverage) and in general, both parameters should be high for variants to be called confidently. On the other hand, the Cameroon target population only contained short regions of the genome chosen based on previous GWAS performed by Doyle et al., (2017) and Hedtke et al. (2017) in Cameroon and on Ghana worms, with a coverage of <1% of the genome. However, the sequences were of high quality, resulting in high depth and coverage of the regions targeted in the reduced representation experiment. In this case, there was no correlation between depth and coverage because the sequencing method (reduced representation) deliberately ensured that only a tiny fraction of the genome was sequenced.

The filtering of the pre-imputation target population with the *beagle* filtering with *conform-gt* tool could also influence the imputation accuracy. The filtering was done to remove variant sites that are not in the reference panel but that there was no filtering of the reference panel to remove variant sites that are not in the target populations. This means that when using a reference that includes individuals drawn from a different population (or breed, in livestock) some reference variants will be imputed even if they are not present at all in the target population, that is, there will be more imputed variants, but they may be at loci that are not polymorphic in the target population or that do not exist in the target population. There was a much larger decrease in the variants retained after filtering the Cameroon target population with the *beagle* filtering with *conform-gt* tool relative to the decrease observed in the W. African target population (10x more variant sites in the W. African target population was observed compared to the Cameroon target population after *beagle* filtering with *conform-gt*). This reason for the difference was because the total genome coverage of the Cameroon target population was missing

99%, while the reference population was either (a) a global reference that was composed mostly of worms from outside Cameroon and could poorly represent the target population; that is, population structure interfered with inference and/or (b) a Cameroon reference panel with few individual sequenced at relatively low depth. Both reasons suggest that there were variants in the Cameroon target population that were not found in the global or the Cameroon reference panels during the *beagle* filtering.

Imputation can be affected by phasing accuracy in the reference panel (Li et al., 2009, Marchini and Howie, 2010), which in turn is very dependent on LD. Inaccuracy in the estimated phase of the reference haplotypes from genotypic data would limit imputation accuracy. Phasing was not possible in the Cameroon data. The reason for that may be because the insufficient SNP density that resulted in low levels of LD, which in turn makes it difficult to call phase. That is, the SNP density (most acutely in the target population, which is composed of reduced representation sequences) is smaller than is required for phasing (from the LD study in previous chapter, SNP density of approximately 1.5kb intervals across the genome is essential for successful genotype imputation and GWAS) (chapter three).

4.4.2. Association studies.

I showed in this study how imputation strengthens the genetic association between some variants with ivermectin response phenotype and reduces it strength for others. This will aid in achieving the broad aim of identifying markers that can be used to develop diagnostic tools in the field.

A reduction in the number of apparently significantly associated variant sites was observed when imputation was done with the W. African reference panel (number of variant sites at the $p<10^{-3}$ threshold = 23) when compared to the pre-imputed variants ($p < 10^{-3}$ n = 25). It may be that there was a better agreement between the reference and imputed data when using a W. African reference, that is, fewer outliers because of population specific differences (and which has also helped to decrease the risk of “noise”). Population structure (from the non-W. African individuals in the global reference) are likely responsible for the increase in the number of “significant” variant sites in the Global/W. Africa imputation combination (number of variant sites at the $p<10^{-3}$ threshold = 74). What imputation did, in the W. African data, was to generally reduce the strength of association of variant sites in regions that showed some association prior to imputation. The reduction in p-value (that is, increase in strength of association) was rather small and well short of what was required to meet the more stringent threshold set by Bonferroni correction. Increasing the size and quality of the reference panel and increasing (if possible) the size of the GR/SOR target might improve this. From a practical perspective, increasing the quality of the reference is more feasible.

The impact of imputation on the Cameroon data was quite different. The pre-imputation GWAS had many variants, distributed along the entire length of the autosomes that show strong associations. The effect of imputation was to remove most of these variants, leaving a much lower "background" level of association with a small number of variant sites that remain above the threshold (albeit at a much-reduced significance level). Imputation helped boost the identification of candidate loci that may then be prioritized for follow-up analysis or analysed in the context of biological pathways. This was clear in the GWAS carried out on the Cameroon population (reduced representation sequences using Nugen Allegro genotyping). For example, the number of markers that showed strongest association (at Bonferroni correction value of $p < 0.05$) with ivermectin resistance dropped from 500 in pre-imputation data to 10 after imputation with the Cameroon and global reference panel.

The outcome of imputation on the Cameroon population suggests that more variant sites that are evenly distributed are needed to be genotyped for imputation (which relies on LD between the sites) and eventual GWAS to be successful. The regions covered in the Cameroon target population were chosen based on previous GWAS (in Cameroon and in Ghana) (Doyle et al., 2017 and Hedtke et al. 2017), and the fundamental assumption was that marker density within those chosen regions would allow detection of association if it existed but that any associations outside of the regions sequenced would not be detected. As a result, the variants were picked unevenly across the genome. While there was a significant increase in the number of sites available in the Cameroon population, because all individuals were uniformly missing approximately 99% of the genome, it was not practical (or reasonable to expect) that imputation could fill in gaps of that magnitude. Increasing the size and quality of the reference panel and the size of the GR/SOR target could be the next step to consider but with the added requirement (if possible) of increasing the coverage of the target population genome by increasing marker density with an expanded reduced representation panel (assuming more DNA can be gotten from the GR/SOR populations).

To carry out any association study, quantifying the extent of LD is vital to determining the number of markers required to associate genetic variation with a phenotype with concise power and precision (Meadows et al., 2008). In the previous chapter (chapter three; section 3.3), I reported that average LD is low in the *O. volvulus* genome (average LD between adjacent SNP loci = 0.25 and 0.21 for chromosomes 1 and 4 respectively) and LD breaks down rapidly (< 1.5kb) as we move away from a QTL. The point at which LD falls below $r^2 > 0.33$ is the critical value for determining how close a variant site must be to a causative polymorphism before a significant genetic association between the phenotype and the marker will occur. Thus, based on LD decline moving away from a local LD maximum, that is, an elevation of LD, the genome requires a dense marker density spaced by approximately 1.5kb across the genome to localize an association between marker and phenotype.

Conclusion

This is the first study to provide a conceptual insight into genomic imputation in helminth parasite. Genotype imputation is a novel approach to improving the power of GWAS to detect genetic association in filarial nematodes. Even though imputation could be a powerful tool for minimizing costs associated with genetic-based screening for sub-optimal response in *O. volvulus* or for drug resistance in other helminths, it has not been tested previously in helminth parasites mainly because of the challenges associated with performing GWAS. These challenges include small sample size, cost of genome sequencing at high depth, the low availability of quality assembled genome to use as reference for variant calling, and the lack of high depth genome sequences that could be used as reference panels for imputation.

Having access to a relatively (for helminths) large repository of unpublished genome sequences for *O. volvulus* availed me the opportunity to test the benefits and limitations of genomic imputation in *O. volvulus*. Thus, I have established in this chapter a conceptual study on the feasibility and accuracy of imputation with regard to the appropriate reference panel to use and the success of imputation in improving the power of association of variants with drug resistance.

Imputation is a research tool that should increase the ability to define a panel of SNPs that are predictive of drug response. That panel would be the basis of a genotyping tool that could be deployed in endemic countries, for example, a PCR-based genotyping diagnostic, or a genotyping LAMP assay etc. I have shown how considerations of reference panel origin and quality, and target population genotype data, might affect the design of an imputation experiment in helminths. There is a question of how widely applicable imputation is likely to be in endemic countries because of the impediment of (a) cost of generating adequate target population genotype data for statistically robust imputation to be feasible and (b) availability of researchers with the necessary skills and access to suitable computing facilities.

Imputation provides researchers with an alternative to amplicon resequencing. In the current study, I have used reference panels sequenced at a greater depth to impute low-coverage and partially sequenced samples. However, one important conclusion from this study is that the choice and quality of the reference panel has a substantial impact on imputation accuracy and a much broader effect on probabilities of association for GWAS.

A reference population that takes population structure into account is required to help improve the power for association studies.

The genotyping method used to generate data for target populations is also essential to consider. For example, the reduced representation sequencing of approximately 10,000 short targets (that generated >10,000 variant sites and is analogous to amplicon resequencing) did not produce sufficiently dense genotype data in a target population for imputation of the entire genome owing to the genotyping technology used. This, combined with chapter three data that showed that LD declines below useful levels in less than 1.5kb suggests that the only feasible genotyping platforms to generate suitable target populations at this stage are low coverage whole genome sequencing or a SNPchip of perhaps 100,000 markers (which would still have an average spacing of approximately 1 – 2 kb). I recommend mapping of the imputed variants in the Cameroon target population having AR2>0.8 to the genome. This could help to reveal what proportion of them mapped to the regions covered by the Allegro panel because most of those higher quality imputed Cameroon variants will likely be close to/within the targeted Allegro panel. This will further help in confirming the appropriate spacing between variant sites essential for imputation.

Chapter Five

General Discussion

5.1.**Summary of Results.**

Onchocerca volvulus is a parasitic nematode which causes the disease onchocerciasis (or river blindness) in humans and is transmitted by repeated bites of infected *Simulium* spp. blackflies. Millions of people are infected with onchocerciasis worldwide, of which the majority lives in Africa (Centers for Disease Control, 2019). The microfilariae cause the major symptoms of onchocerciasis while navigating through the body of the human host and after their death. The major tool used to tackle the menace of *O. volvulus* is mass drug administration (MDA) with ivermectin (a macrocyclic lactone (ML) broad spectrum anthelmintic) (World Health Organization, 2019). The drug clears microfilariae from the skin of infected people and temporarily suppresses the production of new microfilariae by the adult parasite (Duke et al., 1991). However, the repeated use of ivermectin as a preventive chemotherapy has resulted in the emergence of sub-optimal response (SOR) in central Ghana and in the Mbam and Nkam valleys in Cameroon (Awadzi et al., 2004, Bourguinat et al., 2007, Nana-Djeunga et al., 2012, Osei-Atweneboana et al., 2007). This poses a threat to the long-term elimination goals of MDA programs in those foci (Dadzie et al., 2003).

As mentioned earlier, SOR has only been documented in two foci: central Ghana and in the Mbam and Nkam valleys in Cameroon. However, it may be more widespread, but no one is looking and there are no reports of SOR other than those two foci. This is a major concern and is in fact one of the major motivations for this project. It is clear that the phenotypic detection via embryogram is completely impractical (Churcher et al., 2009, Osei-Atweneboana et al., 2011) and the impracticality of the assay is a primary reason for the failure of control programs to institute any form of resistance surveillance (Dadzie et al., 2003). To face this growing threat, there is the need for a genotypic assay based on an understanding of the genetic basis of SOR. My work is an important step towards developing the genotypic assay that may be more feasible to apply on the scale required. I tested the hypothesis that sub-optimal response to ivermectin in *O. volvulus* is genetically determined, such that LD will exist around ivermectin-response loci (or quantitative trait loci (QTLs)) that are under selection and will allow rational experimental design for

GWAS of *O. volvulus*. In achieving this, I developed a methodological framework that could help in achieving the broad aim of my project which is to improve diagnostic capability of SOR in *O. volvulus* and aid elimination goals. This methodological framework encompassed amplicon resequencing to validate putative QTLs, identification of LD structure between QTLs and causative SNPs and genotype imputation to increase the visibility of the causative SNPs or loci.

To develop a genotypic assay, one needs validated genetic markers predictive of ivermectin response. There are two approaches to the discovery of those markers: candidate genes and *de novo* GWAS (Nana-Djeunga et al., 2012, Osei-Atweneboana et al., 2012, Doyle et al., 2017, Hedtke et al., 2017). The candidate gene approach involves carrying out analyses on genes chosen based on specific hypotheses concerning mechanisms of resistance to the drug. Candidate gene studies in ivermectin-resistance associated test identified several different candidates that may be involved but the problem with candidate gene studies in *O. volvulus* is the poor quality of the statistical testing which fails to take population structure and multiple testing correction into account. That implies that none of the studied candidate genes to date met the standard of validation (Doyle and Cotton, 2019). Also, other experiments claiming associations are poorly designed and the data produced have been interpreted incorrectly (Doyle and Cotton, 2019, Hedtke et al., 2019). There is no problem with a candidate gene approach in establishing the mechanism of antihelmintics resistance in nematodes if one has detailed information on mechanism from some other approach. For example, in the mechanism of benzimidazole (BZ) resistance in nematodes, biochemical experiments showed clearly that (i) BZ drugs bind to specific sites on beta-tubulin (ii) that the nematode specificity of BZ-antihelmintics was correlated with the affinity of the drug for nematode beta-tubulin (iii) that BZ binding affinity was reduced in BZ-resistant nematodes (Kwa et al., 1993, Lacey and Gill, 1994, Grant and Mascord, 1996, Lacey and Snowdon, 1988). Given the lowered affinity for BZ of tubulin from resistant worms in particular, it was perfectly reasonable to hypothesise that polymorphism in the beta-tubulin gene(s) would likely be associated with response in nematodes. This was shown to be the case, and the causal role of the mutations was also established by transgenesis in *C. elegans* (Roos et al., 1995). However, no such body of data exist for ML's, and candidate gene studies in ML's are flawed because those independent sources of support for a causal relationship are missing. For example, in *H. contortus*, Gill et al. (1991) showed that there was no pharmacological difference between ivermectin-sensitive and ivermectin-resistant strains of *H. contortus*. In other words, there is biochemical evidence that resistance is not associated with receptor pharmacology in the parasite (Gill et al., 1991). From subsequent association studies, selection for ML resistance was suggested to emerge from a more complex genetic mechanisms, that is, it involves soft selection on multiple quantitative QTLs (Bourguinat et al., 2015, Choi et al., 2017, Doyle et al., 2017, Hedtke et al. 2017). As a result, genome-wide scan replaced candidate gene approaches in association studies of ML resistance in helminth parasites because genome scan gets around the problem of detecting and describing multiple genes involved in soft selection and QTLs (Gilleard, 2006).

Genome-wide scan was used successfully to identify multiple loci (QTLs) responsible for ML resistance in *D. immitis* (Bourguinat et al., 2015) and *T. circumcincta* (Choi et al., 2017). Similarly, it was used to identify QTLs that are under selection for ivermectin response in *O. volvulus* (Doyle et al., 2017, Hedtke et al., 2017). The important feature of these genome-wide scans is that only one of the many proposed candidate genes were detected (nothing in Bourguinat, et al., 2015, Tcir-pgp-9 in Choi et al., 2017, and no candidate genes in Doyle et al., 2017 and Hedtke et al., 2017). So, in the four genome-wide scans, only one candidate gene confirmed for ML-resistance. It is important to note that genome-wide scan is far from perfect, especially when using limited population set, poorly defined phenotypes and improving genomic resources. Both approaches are fine and needed to understand how a drug works (candidate gene approach) and how selection for resistance emerges under natural selection (QTL study). The only unacceptable issue is to claim that one receptor defines the resistance development process.

Doyle et al.'s genome-wide scan of SOR in *O. volvulus* was based on limited number of low sequence coverage Pool-seq worms which resulted into a stochastic variation in allele detection in the worms. Variants that differentiated GR and SOR parasites were found in several QTLs, but additional studies were required, including examining single whole genome sequences to validate those QTLs (Doyle et al., 2017). Although Pool seq is more economical than single whole genome sequencing, it leads to uncertainty in estimating allele frequency and loss of haplotype information. Hedtke et al.'s genome-wide of SOR in *O. volvulus* was based on sequenced whole-genome single worms from the same study sites, putative QTLs were identified and there was further need to validate identified

QTLs on larger sample size (Hedtke et al., 2017). There are serious technical barriers to the routine use of whole genome sequencing of embryogram phenotyped adult females *O. volvulus*. The technical barriers include impracticality of the phenotype. This was precipitated by the process by which the phenotyped worms were accessed – palpable nodules surrounding adult worms were surgically removed from the infected person (Richards et al., 2000); excised nodules were digested to isolate adult worms and embryograms were prepared with females for the evaluation of their reproductive capacities (Nana-Djeunga et al., 2014, Osei-Atweneboana et al., 2011); the female worms were further categorised into phenotypes based on the presence of stretched microfilariae in their embryogrammes after 80 - 90 days of ivermectin treatment - this process is random with respect to the worm(s) that are removed, which implies that it is impossible to ensure that the worms that were phenotyped were responsible for the microfilariae in the skin. This further increases the sample size that is required for successful GWAS. Limited sample size was also caused by poor quality and low concentration of genomic DNA that can be prepared from worms isolated under field conditions in developing countries. Whole genome sequencing and genotyping of all the identified variants in the genome is potentially expensive and almost certainly impractical (given the DNA quality and/or quantity, except for recent improved technology like optimized amplification), this limits the number of whole genomes sequenced single worm available for GWAS. Owing to the same reason, many of the worms could not be sequenced and the sequenced worms were of uneven coverage across the genome with lots of missing data.

Given this context, amplicon sequencing rather than whole genome sequencing may be a solution because much smaller amounts of genomic DNA are required and, if amplicons are short, poor DNA quality is less limiting. This was discussed in detail in chapter two. I tested the primary hypothesis that some selected non-synonymous SNP loci that fell within some QTL loci (defined by GWAS) will be predictive of ivermectin response (that is, to validate genetic markers predictive of ivermectin response) and a secondary hypothesis that amplicon re-sequencing allows sample size to be increased. The aim of the chapter was to determine the extent of genetic association between ivermectin-response phenotype and the genotype at non-synonymous SNP loci within QTLs that have strong support from the previous GWAS, using an amplicon resequencing approach and increased sample size. I chose to test the association of the selected SNP loci because they were (i) in a QTL from the GWAS and (ii) the alternative alleles are non-synonymous. However, the data fail to support an association for these loci but do point to (i) selection that is not associated with ivermectin response and (ii) other, novel SNP loci that may be associated with ivermectin response. The new data from a larger sample size did not support a role for those SNP loci because of many reasons that had already been discussed in chapter two of this thesis. The secondary hypothesis tested which was *amplicon re-sequencing allows sample size to be increased* was supported because I was able to genotype a larger number of worms, most of which could not have been genotyped by whole genome sequencing because of DNA concentration and/or quality, but it was not possible to draw any conclusion on this hypothesis because that increased sample size did not help in the sense that there was no association detected between those loci and ivermectin response. This result does not imply that the loci/SNPs conferring resistance do not exist, rather, there was a need for refinement of QTLs before further validation tests are carried out on them. The hypothesis of association between a SNP locus and the phenotype requires either that the SNP locus itself is causal or that it is in LD with a causal locus. The chosen loci are not causal, but they do fall within a broad QTL locus. The discrepancy may be because either the QTL locus is an artefact of the GWAS, or the locus might be real, but the SNP loci tested might not be in strong LD with the causal locus/loci within the QTL. Which point to the fact that there is a need to understand the LD structure in the genome as a whole and around putative QTLs in particular.

In chapter three, I characterised LD and most importantly, defined haploblocks and thresholds for detection of an association between a SNP locus and a causal mutation. This availed me the opportunity to predict exactly how many SNP loci at what sort of density are required for GWAS. Since genotypes at nearby markers are usually correlated (that is, they are in LD), it may be possible to scan the genome using a much smaller marker set with only a modest loss of power to detect selection while minimising the quantity and quality of DNA that is required. Studies of LD in nematodes has only been limited to LD study across six short nuclear loci of wild isolates of the gonochoristic *C. remanei*. It was suggested that LD declines significantly over just a few hundred base pairs at a rate suggesting that linkage equilibrium will be reached at distances of 1–2 kb (Cutter et al., 2006). Similarly, measures of LD in 96 *O. volvulus* samples around a QTL suggested that LD decays rapidly in *O.* volvulus and reaches the threshold for useful LD estimate at an approximate distance of 1.5kb. In the same vein, haploblock structures in

O. volvulus genome was characterised by clusters of small, fragmented blocks of low to moderately elevated LD that correlate with peaks of F_{ST} (or putative QTL). Based on the extent of LD up to the value of useful LD (that is, up to the point where $r^2 = 0.33$) as we move away from a QTL, the genome requires a dense SNP spaced by approximately 1.5kb distance across the genome and a minimum SNP density of 20,775 and 10,699 are needed in OM1 and OM4 chromosomes, respectively, to confidently detect the association of a SNP loci with a trait of interest in *O. volvulus.* Applying this principle to the entire *O. volvulus* genome, approximately 64,668 fully informative SNPs are needed to saturate the entire *O. volvulus* genome to confidently detect the association of a SNP loci with a trait of interest. This is still a big number, and that led me to the hypothesis that it may be possible to impute SNP loci at this density from a smaller genotyping panel as has been done successfully for a range of other species (humans, cows, sheep, several crop plants, etc) (Cavalli-Sforza, 2005, Daetwyler et al., 2014, Ventura et al., 2016).

In Chapter four, I explored the feasibility and accuracy of genotype imputation by making use of two different sets of reference panel and test the success of imputation in improving the power of association of SNPs with ivermectin response. Thus, mitigating the difficulty of sufficient sampling for GWAS. Imputation is an *in-silico* method that can increase the power of association studies by inferring missing/un-typed genotypes (VanRaden et al., 2013). Imputation is a new tool in filarial nematode population genetic studies, and reference panels for imputing missing genotypes have not been previously developed in them. Imputation could be a powerful tool for minimizing the costs associated with genetic-based screening for sub-optimal response in *O. volvulus* or for drug resistance in other helminths. To design imputation studies, it is necessary to have a detailed understanding of the structure and extent of LD across the genome, both to choose suitable reference and genotyping marker sets (Li et al., 2009). Having established the structure and the extent of useful LD in *O. volvulus* in chapter three and having access to largest single worm's whole genome sequences for *O. volvulus,* I maximized the chance to test the benefits and limitations of genotype imputation in *O. volvulus*. I was able to establish in this thesis a conceptual study on the feasibility and accuracy of imputation regarding the appropriate reference panel to use and the success of imputation in improving the power of association of SNP loci with drug resistance. I showed that imputation is likely feasible from low depth whole genome sequence but probably not feasible from reduced representation data such as Nugen Allegro, which does not have sufficient SNP density and spacing because of the low LD in the *O. volvulus* genome overall. My strong recommendation for improving the challenge of low imputation quality from reduced representation data is to map the relatively small number of the high quality (AR2>0.8) imputed variants from the reduced representation data to the genome in order to test the hypothesis that those high-quality imputed Cameroon variants will be close to or within the NuGen targeted regions of the genome. That will aid in identifying the appropriate reduced representation panel to be designed and at what SNP spacing that will be sufficient for imputation and GWAS

5.2. **Conclusions and Future Prospects.**

Going back to the original problem which is how to develop a genotypic assay predictive of ivermectin response. Overall, I have been a been able to estimate LD, which is a major criterion in predicting the genetic marker density that is required for carrying out successful GWAS in *O. volvulus* and was able to show that imputation of SNP loci at that density from a smaller genotyping panel can be done successfully. The major outcomes of this study are the clear understanding provided about the density, spacing, and number of useful SNPs essential for imputation and GWAS, and identification of the density of SNPs needed for designing a SNP array for future use in developing diagnostic tools. Therefore, it is easy to conclude that the prospects are better now that we know what genetic marker density is required for GWAS. This will aid in the end goal of eliminating onchocerciasis in Africa.

The first steps towards the conceptual study to test the feasibility and accuracy of imputation in *O. volvulus* and to test the success of imputation in improving the power of association of SNP loci with drug resistance was successful. Therefore, I recommend that genotype imputation should be implemented in helminths because it could be a powerful tool for minimizing costs associated with genetic-based screening for drug resistance in helminths. Although, broadening of imputation to other helminths is subjected to many issues that should be considered, including, its limited application for more diverse parasitic populations like *H. contortus.*

I have shown how considerations of reference panel origin, quality and sequencing depth, and target population genotype data, might affect the design of an imputation experiment in helminths. From this study, I observed that the choice and quality of the reference panel has a substantial impact on imputation accuracy and a much broader effect on probabilities of association for GWAS. Therefore, I recommend using a reference panel derived from the same population as the target data as a better option than a geographically diverse reference because of population structure. Imputation from a genetically divergent population in which haplotype frequencies may differ will result in lower confidence imputation, or imputation of haplotypes that may be absent in the target population.

I have been able to measure LD, which is a major criterion in knowing the genetic marker density that is required for successful GWAS in *O. volvulus* and I have been able to show that imputing SNPs at that density from a smaller genotyping panel can be done successfully. However, the problem of sample size is not resolved entirely. The problem of sample size still remains in establishing this methodological framework because genotype imputation as a tool needs a large reference panel. The sample size problem for imputation is majorly the size of the reference panel: large numbers of samples (>100 individuals) are usually used in human and cattle, where imputation is routinely used (Anderson et al., 2008, Daetwyler, 2020, Korkuc et al., 2019). Small reference panels for imputation limits imputation accuracy while larger reference panels substantially increase imputation accuracy, particularly for low-frequency variants (Browning and Browning, 2009). The requirements for a reference panel based on the imputation results observed in this study are larger samples size, a reference panel population as the same with the target population (that is population structure must be considered). Therefore, next step in the research program should identify population structure to define how many samples are required as reference panel. Having in mind that larger reference panels require large-scale sequencing and genotyping projects, better sequencing methods should be developed so that these reference panels can be assembled using sequencing from microfilariae (which are readily available in large quantities) rather than adult worms.

I recommend further evaluation studies on the 20 genetic associations between specific variant site and ivermectin response phenotype that met the Bonferroni correction of $p < 0.05$ threshold in the Cameroon target samples imputed with the global and Cameroon reference panel respectively. This can be a step further in achieving the broad aim of developing genotyping assay for field diagnosis of SOR.

I strongly recommend that new GWAS to be performed on imputed variants should consider using a program that can make use of dosage data - this helps model the imputation uncertainty. Imputation uncertainty was considered in this study by manually eliminating variant position with low imputation accuracy (AR2 < 0.8). However, plink 2.0 can make use of dosage data in association studies and can be used henceforth.

Finally, in this study, amplicon resequencing and LD study did not validate the potential QTL for SOR resistance in *O. volvulus*. However, there are some other interesting results that emerged from the study which are not related with drug resistance in the worm: I observed regions of elevated LD in regions showing low differentiation between GR and SOR (that is, higher LD and bigger haploblocks in SOR worms and in region B in chapter three, results section). The nature of this selection is not known, but a reasonable candidate might be the host switch that gave rise to *O. volvulus* (Oncho speciation) or could simply be a region of higher gene flow. It is not surprising, but it is the first time this has been done in *O. volvulus*. The expectation was that strong LD will occur with recent selection, but the opposite was observed. Although selection for drug resistance is a soft selection but not weak, that is, lots of QTLs caused the phenotype. These QTLs cannot be found in the SOR worms at once and does not necessarily infer that they are solely causing SOR in the worm. These further enriches the evidence of soft selection and is novel in the genomics of this parasite. This suggests that new strategies for validation of QTL's can be implemented, for example, AmpliSeq panels, which is a genotyping method that aims to generate data that can be used for GWAS or for targeting a large number of SNP loci (as used, for example, in cancer diagnostic tests).

Bibliography

1000 GENOMES PROJECT CONSORTIUM 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

AKOGUN, O., AKOGUN, M. & AUDU, Z. 2000. Community-perceived benefits of ivermectin treatment in northeastern Nigeria. *Social science & medicine,* 50**,** 1451-1456.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-410.

ANDERSON, C. A., PETTERSSON, F. H., BARRETT, J. C., ZHUANG, J. J., RAGOUSSIS, J., CARDON, L. R. & MORRIS, A. P. 2008. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *The American Journal of Human Genetics,* 83**,** 112-119.

ANDERSON, S. L., MAHAN, A. L., MURRAY, S. C. & KLEIN, P. E. 2018. Four parent maize (FPM) population: Effects of mating designs on linkage disequilibrium and mapping quantitative traits. *The plant genome,* 11.

ANDRADE, A. C. B., VIANA, J. M. S., PEREIRA, H. D., PINTO, V. B. & E SILVA, F. F. 2019. Linkage disequilibrium and haplotype block patterns in popcorn populations. *PloS one,* 14.

ANGIUS, A., HYLAND, F. C., PERSICO, I., PIRASTU, N., WOODAGE, T., PIRASTU, M. & FRANCISCO, M. 2008. Patterns of linkage disequilibrium between SNPs in a Sardinian population isolate and the selection of markers for association studies. *Human Heredity,* 65**,** 9-22.

ARCHIBALD, A., COCKETT, N., DALRYMPLE, B., FARAUT, T., KIJAS, J., MADDOX, J., MCEWAN, J., HUTTON ODDY, V. & RAADSMA, H. 2010. The sheep genome reference sequence: a work in progress. *Animal genetics,* 41**,** 449-453.

ARDELLI, B. F., GUERRIERO, S. B. & PRICHARD, R. K. 2005. Genomic organization and effects of ivermectin selection on *Onchocerca volvulus* P-glycoprotein. *Molecular and biochemical parasitology,* 143**,** 58-66.

ARDELLI, B., GUERRIERO, S. & PRICHARD, R. 2006. Ivermectin imposes selection pressure on P-glycoprotein from *Onchocerca volvulus*: linkage disequilibrium and genotype diversity. *Parasitology,* 132**,** 375.

ARDLIE, K. G., KRUGLYAK, L. & SEIELSTAD, M. 2002. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics,* 3**,** 299-309.

ARMOO, S., DOYLE, S. R., OSEI-ATWENEBOANA, M. Y. & GRANT, W. N. 2017. Significant heterogeneity in Wolbachia copy number within and between populations of *Onchocerca volvulus*. *Parasites & vectors,* 10**,** 188.

ARNHEIM, N. 1983. Concerted evolution of multigene families. *Evolution of genes and proteins*.

AWADZI, K. 1993. Drug surveillance: international cooperation past, present and future. *Proceedings of the XXVIIth CIOMS (Council for International Organisation of Medical Sciences) Confererence, Geneva***,** 14-15.

AWADZI, K., ATTAH, S. K., ADDY, E. T., OPOKU, N. O., QUARTEY, B. T., LAZDINS-HELDS, J. K., AHMED, K., BOATIN, B. A., BOAKYE, D. A. & EDWARDS, G. 2004b. Thirty-month follow-up of sub-optimal responders to multiple treatments with ivermectin, in two onchocerciasis-endemic foci in Ghana. *Annals of Tropical Medicine & Parasitology,* 98**,** 359-370.

AWADZI, K., ATTAH, S., ADDY, E., OPOKU, N., QUARTEY, B., LAZDINS-HELDS, J., AHMED, K., BOATIN, B., BOAKYE, D. & EDWARDS, G. 2004a. Thirty-month follow-up of sub-optimal responders to multiple treatments with ivermectin, in two onchocerciasis-endemic foci in Ghana. *Annals of Tropical Medicine & Parasitology,* 98**,** 359-370.

AWADZI, K., ATTAH, S., ADDY, E., OPOKU, N., QUARTEY, B., LAZDINS-HELDS, J., AHMED, K., BOATIN, B., BOAKYE, D. & EDWARDS, G. 2004. Thirty-month follow-up of sub-optimal responders to multiple treatments with ivermectin, in two onchocerciasis-endemic foci in Ghana. *Annals of Tropical Medicine & Parasitology,* 98**,** 359-370.

AWADZI, K., BOAKYE, D. A., EDWARDS, G., OPOKU, N. O., ATTAH, S. K., OSEI-ATWENEBOANA, M. Y., LAZDINS-HELDS, J. K., ARDREY, A. E., ADDY, E. T., QUARTEY, B. T., AHMED, K., BOATIN, B. A. & SOUMBEY-ALLEY, E. W. 2004c. An investigation of persistent microfilaridermias despite multiple treatments with ivermectin, in two onchocerciasis-endemic foci in Ghana. *Annals of Tropical Medicine & Parasitology,* 98**,** 231-249.

BASÁÑEZ, M.G., CHURCHER, T.S. AND GRILLET, M.E., 2009. Onchocerca–Simulium interactions and the population and evolutionary biology of *Onchocerca volvulus. Advances in parasitology*, 68, pp.263-313.

BHATIA, G., PATTERSON, N., PASANIUC, B., ZAITLEN, N., GENOVESE, G., POLLACK, S., MALLICK, S., MYERS, S., TANDON, A. & SPENCER, C. 2011. Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *The American Journal of Human Genetics,* 89**,** 368-381.

BISWAS, S. & AKEY, J. M. 2006. Genomic insights into positive selection. *TRENDS in Genetics,* 22**,** 437-446.

BLACKHALL, W. J., LIU, H. Y., XU, M., PRICHARD, R. K. & BEECH, R. N. 1998. Selection at a P-glycoprotein gene in ivermectin-and moxidectin-selected strains of Haemonchus contortus. *Molecular and biochemical parasitology,* 95**,** 193-201.

BLAGBURN, B., DILLON, A., ARTHER, R., BUTLER, J. & NEWTON, J. 2011. Comparative efficacy of four commercially available heartworm preventive products against the MP3 laboratory strain of *Dirofilaria immitis*. *Veterinary Parasitology,* 176**,** 189-194.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BOTTOMLEY, C., ISHAM, V., VIVAS-MARTÍNEZ, S., KUESEL, A. C., ATTAH, S. K., OPOKU, N. O., LUSTIGMAN, S., WALKER, M. & BASÁÑEZ, M.-G. 2016. Modelling neglected tropical diseases diagnostics: the sensitivity of skin snips for *Onchocerca volvulus* in near elimination and surveillance settings. *Parasites & vectors,* 9**,** 343.

BOURGUINAT, C., ARDELLI, B. F., PION, S. D., KAMGNO, J., GARDON, J., DUKE, B. O., BOUSSINESQ, M. & PRICHARD, R. K. 2008. P-glycoprotein-like protein, a possible genetic marker for ivermectin resistance selection in *Onchocerca volvulus*. *Molecular and biochemical parasitology,* 158**,** 101-111.

BOURGUINAT, C., LEE, A. C., LIZUNDIA, R., BLAGBURN, B. L., LIOTTA, J. L., KRAUS, M. S., KELLER, K., EPE, C., LETOURNEAU, L. & KLEINMAN, C. L. 2015. Macrocyclic lactone resistance in *Dirofilaria immitis*: Failure of heartworm preventives and investigation of genetic markers for resistance. *Veterinary parasitology,* 210**,** 167-178.

BOURGUINAT, C., PION, S. D., KAMGNO, J., GARDON, J., DUKE, B. O., BOUSSINESQ, M. & PRICHARD, R. K. 2007. Genetic selection of low fertile *Onchocerca volvulus* by ivermectin treatment. *PLoS neglected tropical diseases,* 1**,** e72.

BOUSSINESQ, M., CHIPPAUX, J.-P., ERNOULD, J., QUILLEVERE, D. & PROD'HON, J. 1995. Effect of repeated treatments with ivermectin on the incidence of onchocerciasis in northern Cameroon. *The American journal of tropical medicine and hygiene,* 53**,** 63-67.

BOUSSINESQ, M., PROD'HON, J. & CHIPPAUX, J.-P. 1997. *Onchocerca volvulus*: striking decrease in transmission in the Vina valley (Cameroon) after eight annual large scale ivermectin treatments. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 91**,** 82-86.

BOWMAN, D. D. 2012. Heartworms, macrocyclic lactones, and the specter of resistance to prevention in the United States. *Parasites & vectors,* 5**,** 138.

BRADY, S. C., ZDRALJEVIC, S., BISAGA, K. W., TANNY, R. E., COOK, D. E., LEE, D., WANG, Y. & ANDERSEN, E. C. 2019. A novel gene underlies Bleomycin-response variation in *Caenorhabditis elegans*. *Genetics,* 212**,** 1453-1468.

BROWNING, B. L. & BROWNING, S. R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics,* 84**,** 210-223.

BROWNING, B. L. & BROWNING, S. R. 2016. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics,* 98**,** 116-126.

BROWNING, B. L., ZHOU, Y. & BROWNING, S. R. 2018. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics,* 103**,** 338-348.

BROWNING, S. R. & BROWNING, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics,* 81**,** 1084-1097.

BURNHAM, G., 2007. Efficacy of ivermectin against Onchocerca volvulus in Ghana. The Lancet, 370(9593), p.1125.

C. ELEGANS SEQUENCING CONSORTIUM* 1998. Genome sequence of the nematode *C. elegans:* a platform for investigating biology. *Science,* 282**,** 2012-2018.

CARLSON, C. S., EBERLE, M. A., RIEDER, M. J., YI, Q., KRUGLYAK, L. & NICKERSON, D. A. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics,* 74**,** 106-120.

CASEY, P. J. & GILMAN, A. G. 1988. G protein involvement in receptor-effector coupling. *Journal of Biological Chemistry,* 263**,** 2577-2580.

CAVALLI-SFORZA, L. L. 2005. The human genome diversity project: past, present and future. *Nature Reviews Genetics,* 6**,** 333-340.

CENTERS FOR DISEASE CONTROL 2016. Onchocerciasis.

CENTERS FOR DISEASE CONTROL. 2019. *Parasites - Onchocerciasis (also known as River Blindness).* [Online]. Global Health, Division of Parasitic Diseases. [Accessed 7 August 2020].

CHALASANI, G. 2020. *RE: Cameroon reference panels.*

CHENG, S., FOCKLER, C., BARNES, W. M. & HIGUCHI, R. 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences,* 91**,** 5695-5699.

CHESNAIS, C. B., NANA-DJEUNGA, H. C., NJAMNSHI, A. K., LENOU-NANGA, C. G., BOULLÉ, C., BISSEK, A.-C. Z.-K., KAMGNO, J., COLEBUNDERS, R. & BOUSSINESQ, M. 2018. The temporal relationship between onchocerciasis and epilepsy: a population-based cohort study. *The Lancet Infectious Diseases,* 18**,** 1278-1286.

CHOI, Y.-J., BISSET, S. A., DOYLE, S. R., HALLSWORTH-PEPIN, K., MARTIN, J., GRANT, W. N. & MITREVA, M. 2017. Genomic introgression mapping of field-derived multiple-anthelmintic resistance in *Teladorsagia circumcincta*. *PLoS genetics,* 13**,** e1006857.

CHOI, Y.-J., TYAGI, R., MCNULTY, S. N., ROSA, B. A., OZERSKY, P., MARTIN, J., HALLSWORTH-PEPIN, K., UNNASCH, T. R., NORICE, C. T., NUTMAN, T. B., WEIL, G. J., FISCHER, P. U. & MITREVA, M. 2016. Genomic diversity in *Onchocerca volvulus* and its Wolbachia endosymbiont. 2**,** 16207.

CHURCHER, T. S., PION, S. D., OSEI-ATWENEBOANA, M. Y., PRICHARD, R. K., AWADZI, K., BOUSSINESQ, M., COLLINS, R. C., WHITWORTH, J. A. & BASÁÑEZ, M.-G. 2009. Identifying sub-optimal responses to ivermectin in the treatment of River Blindness. *Proceedings of the National Academy of Sciences,* 106**,** 16716-16721.

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly,* 6**,** 80-92.

CLARK, A. G., WEISS, K. M., NICKERSON, D. A., TAYLOR, S. L., BUCHANAN, A., STENGÅRD, J., SALOMAA, V., VARTIAINEN, E., PEROLA, M. & BOERWINKLE, E. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *The American Journal of Human Genetics,* 63**,** 595-612.

CLAUDIANOS, C., RUSSELL, R. J. & OAKESHOTT, J. G. 1999. The same amino acid substitution in orthologous esterases confers organophosphate resistance on the house fly and a blowfly. *Insect biochemistry and molecular biology,* 29**,** 675-686.

COLOSIMO, P. F., HOSEMANN, K. E., BALABHADRA, S., VILLARREAL, G., DICKSON, M., GRIMWOOD, J., SCHMUTZ, J., MYERS, R. M., SCHLUTER, D. & KINGSLEY, D. M. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *science,* 307**,** 1928-1933.

CONNOLLY, S. & HERON, E. A. 2015. Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. *Briefings in Bioinformatics,* 16**,** 429-448.

COOK, D. E., ZDRALJEVIC, S., ROBERTS, J. P. & ANDERSEN, E. C. 2017. CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic acids research,* 45**,** D650-D657.

CORSI, A. K., WIGHTMAN, B. & CHALFIE, M. 2015. A transparent window into biology: a primer on *Caenorhabditis elegans*. *Genetics,* 200**,** 387-407.

COTTON, J. A., BENNURU, S., GROTE, A., HARSHA, B., TRACEY, A., BEECH, R., DOYLE, S. R., DUNN, M., HOTOPP, J. C. D. & HOLROYD, N. 2016. The genome of *Onchocerca volvulus*, agent of river blindness. *Nature microbiology,* 2**,** 16216.

CRAINEY, J. L., DA SILVA, T. R., ENCINAS, F., MARÍN, M. A., VICENTE, A. C. P. & LUZ, S. L. 2016. The mitogenome of *Onchocerca volvulus* from the Brazilian Amazonia focus. *Memórias do Instituto Oswaldo Cruz,* 111**,** 79-81.

CRAWFORD, K. E., HEDTKE, S. M., DOYLE, S. R., KUESEL, A. C., ARMOO, S., OSEI-ATWENEBOANA, M. & GRANT, W. N. 2019. Utility of the *Onchocerca volvulus* mitochondrial genome for delineation of parasite transmission zones. *bioRxiv***,** 732446.

CULLY, D. F., VASSILATIS, D. K., LIU, K. K., PARESS, P. S., VAN DER PLOEG, L. H., SCHAEFFER, J. M. & ARENA, J. P. 1994. Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*. *Nature,* 371**,** 707-711.

CUPP E, RICHARDS F, LAMMIE P, EBERHARD M. 2007. Efficacy of ivermectin against Onchocerca volvulus in Ghana. Lancet. 2007; 70(9593):1123; author reply 4-5.

CUTTER, A. D., BAIRD, S. E. & CHARLESWORTH, D. 2006. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics,* 174**,** 901-913.

DABORN, P. J. & LE GOFF, G. 2004. The genetics and genomics of insecticide resistance. *TRENDS in Genetics,* 20**,** 163-170.

DABORN, P., BOUNDY, S., YEN, J. & PITTENDRIGH, B. 2001. DDT resistance in *Drosophila* correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. *Molecular Genetics and Genomics,* 266**,** 556-563.

DADZIE, Y., NEIRA, M. & HOPKINS, D. 2003. Final report of the Conference on the eradicability of Onchocerciasis. *Filaria Journal,* 2**,** 2.

DAETWYLER, H. D. 2020. *RE: Genomic Imputation.*

DAETWYLER, H. D., CAPITAN, A., PAUSCH, H., STOTHARD, P., VAN BINSBERGEN, R., BRØNDUM, R. F., LIAO, X., DJARI, A., RODRIGUEZ, S. C. & GROHS, C. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics,* 46**,** 858.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T. & SHERRY, S. T. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DAS, S., FORER, L., SCHÖNHERR, S., SIDORE, C., LOCKE, A. E., KWONG, A., VRIEZE, S. I., CHEW, E. Y., LEVY, S. & MCGUE, M. 2016. Next-generation genotype imputation service and methods. *Nature genetics,* 48**,** 1284-1287.

DE SOUSA DIAS, M., HERNAN, I., PASCUAL, B., BORRÀS, E., MAÑÉ, B., GAMUNDI, M. J. & CARBALLO, M. 2013. Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing. *Molecular vision,* 19**,** 654.

DENG, T., LIANG, A., LIU, J., HUA, G., YE, T., LIU, S., CAMPANILE, G., PLASTOW, G., ZHANG, C. & WANG, Z. 2019. Genome-Wide SNP Data Revealed the Extent of Linkage Disequilibrium, Persistence of Phase and Effective Population Size in Purebred and Crossbred Buffalo Populations. *Frontiers in genetics,* 9**,** 688.

DENT, J. A., SMITH, M. M., VASSILATIS, D. K. & AVERY, L. 2000. The genetics of ivermectin resistance in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences,* 97**,** 2674-2679.

DOMINGUES, V. S., POH, Y. P., PETERSON, B. K., PENNINGS, P. S., JENSEN, J. D. & HOEKSTRA, H. E. 2012. Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution: International Journal of Organic Evolution,* 66**,** 3209-3223.

DOYLE, S. R. & COTTON, J. A. 2019. Genome-wide approaches to investigate anthelmintic resistance. *Trends in parasitology*.

DOYLE, S. R., BOURGUINAT, C., NANA-DJEUNGA, H. C., KENGNE-OUAFO, J. A., PION, S. D., BOPDA, J., KAMGNO, J., WANJI, S., CHE, H. & KUESEL, A. C. 2017. Genome-wide analysis of ivermectin response by *Onchocerca volvulus* reveals that genetic drift and soft selective sweeps contribute to loss of drug sensitivity. *PLoS neglected tropical diseases,* 11**,** e0005816.

DOYLE, S. R., LAING, R., BARTLEY, D. J., BRITTON, C., CHAUDHRY, U., GILLEARD, J. S., HOLROYD, N., MABLE, B. K., MAITLAND, K. & MORRISON, A. A. 2018. A genome resequencing-based genetic map reveals the recombination landscape of an outbred parasitic nematode in the presence of polyploidy and polyandry. *Genome biology and evolution,* 10**,** 396-409.

DUKE, B.O., 1980. Observations on *Onchocerca volvulus* in experimentally infected chimpanzees. *Tropenmedizin und Parasitologie*, 31(1), pp.41-54.

DUKE, B., ZEA-FLORES, G. & MUNOZ, B. 1991. The embryogenesis of Onchocerca volvulus over the first year after a single dose of ivermectin. *Tropical medicine and parasitology: official organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ),* 42**,** 175-180.

ENG, J. & PRICHARD, R. 2005. A comparison of genetic polymorphism in populations of *Onchocerca volvulus* from untreated-and ivermectin-treated patients. *Molecular and biochemical parasitology,* 142**,** 193-202.

ERTTMANN, K., MEREDITH, S., GREENE, B. & UNNASCH, T. R. 1990. Isolation and characterization of form specific DNA sequences of *O. volvulus. Acta Leidensia,* 59**,** 253-260.

EVANS, K. S. & ANDERSEN, E. C. 2020. The Gene scb-1 Underlies Variation in *Caenorhabditis elegans* Chemotherapeutic Responses. *G3: Genes, Genomes, Genetics,* 10**,** 2353-2364.

FEULNER, P. G., CHAIN, F. J., PANCHAL, M., EIZAGUIRRE, C., KALBE, M., LENZ, T. L., MUNDRY, M., SAMONTE, I. E., STOLL, M. & MILINSKI, M. 2013. Genome‐wide patterns of standing genetic variation in a marine population of three‐spined sticklebacks. *Molecular ecology,* 22**,** 635-649.

FLINT-GARCIA, S. A., THORNSBERRY, J. M. & BUCKLER IV, E. S. 2003. Structure of linkage disequilibrium in plants. *Annual review of plant biology,* 54**,** 357-374.

FREMPONG, K.K., WALKER, M., CHEKE, R.A., TETEVI, E.J., GYAN, E.T., OWUSU, E.O., WILSON, M.D., BOAKYE, D.A., TAYLOR, M.J., BIRITWUM, N.K. AND OSEI-ATWENEBOANA, M., 2016. Does increasing treatment frequency address suboptimal responses to ivermectin for the control and elimination of river blindness? Clinical Infectious Diseases, 62(11), pp.1338-1347.

FU, W., O'CONNOR, T. D., JUN, G., KANG, H. M., ABECASIS, G., LEAL, S. M., GABRIEL, S., RIEDER, M. J., ALTSHULER, D. & SHENDURE, J. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature,* 493**,** 216-220.

GABRIEL, S. B., SCHAFFNER, S. F., NGUYEN, H., MOORE, J. M., ROY, J., BLUMENSTIEL, B., HIGGINS, J., DEFELICE, M., LOCHNER, A. & FAGGART, M. 2002. The structure of haplotype blocks in the human genome. *Science,* 296**,** 2225-2229.

GARCÍA-GÁMEZ, E., SAHANA, G., GUTIÉRREZ-GIL, B. & ARRANZ, J.-J. 2012. Linkage disequilibrium and inbreeding estimation in Spanish Churra sheep. *BMC genetics,* 13**,** 43.

GARDON, J., BOUSSINESQ, M., KAMGNO, J., GARDON-WENDEL, N. & DUKE, B. O. 2002. Effects of standard and high doses of ivermectin on adult worms of *Onchocerca volvulus*: a randomised controlled trial. *The Lancet,* 360**,** 203-210.

GARRISON, E. & MARTH, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

GEARY, T. G. 2005. Ivermectin 20 years on: maturation of a wonder drug. *Trends in parasitology,* 21**,** 530-532.

GILL, J. H., REDWIN, J. M., VAN WYK, J. A. & LACEY, E. 1991. Detection of resistance to ivermectin in *Haemonchus contortus*. *International journal for parasitology*, 21, 771-776.

GILLEARD, J. & BEECH, R. 2007. Population genetics of anthelmintic resistance in parasitic nematodes. *Parasitology,* 134**,** 1133-1147.

GILLEARD, J. S. 2006. Understanding anthelmintic resistance: the need for genomics and genetics. *International journal for parasitology,* 36**,** 1227-1239.

GILLY, A., SOUTHAM, L., SUVEGES, D., KUCHENBAECKER, K., MOORE, R., MELLONI, G. E., HATZIKOTOULAS, K., FARMAKI, A.-E., RITCHIE, G. & SCHWARTZENTRUBER, J. 2019. Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics,* 35**,** 2555-2561.

GODDARD, M. E. & HAYES, B. J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics,* 10**,** 381-391.

GOODE, E. L. 2011. Linkage Disequilibrium. *In:* SCHWAB, M. (ed.) *Encyclopedia of Cancer.* Berlin, Heidelberg: Springer Berlin Heidelberg.

GRANT, W. 2000. What is the real target for ivermectin resistance selection in *Onchocerca volvulus*? *Parasitology today,* 16**,** 458-459.

GRANT, W.N. 2020. *RE: failure of long-range PCR due to DNA quality problem.*

GUDBJARTSSON, D. F., HELGASON, H., GUDJONSSON, S. A., ZINK, F., ODDSON, A., GYLFASON, A., BESENBACHER, S., MAGNUSSON, G., HALLDORSSON, B. V. & HJARTARSON, E. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nature genetics,* 47**,** 435-444.

GUO, Z., HOOD, L., MALKKI, M. & PETERSDORF, E. W. 2006. Long-range multilocus haplotype phasing of the MHC. *Proceedings of the National Academy of Sciences,* 103**,** 6964-6969.

HAASL, R. J., JOHNSON, R. C. & PAYSEUR, B. A. 2014. The effects of microsatellite selection on linked sequence diversity. *Genome biology and evolution,* 6**,** 1843-1861.

HAHNEL, S. R., ZDRALJEVIC, S., RODRIGUEZ, B. C., ZHAO, Y., MCGRATH, P. T. & ANDERSEN, E. C. 2018. Extreme allelic heterogeneity at a *Caenorhabditis elegans* beta-tubulin locus explains natural resistance to benzimidazoles. *PLoS pathogens,* 14**,** e1007226.

HAMPSHIRE, V. A. 2005. Evaluation of efficacy of heartworm preventive products at the FDA. *Veterinary parasitology,* 133**,** 191-195.

HARTL, D. & CLARK, A. 2007. *Principles of population genetics.,* Sunderland, Massachusetts, Sinauer Associates, Inc. Publishers.

HAUSER, A. S., CHAVALI, S., MASUHO, I., JAHN, L. J., MARTEMYANOV, K. A., GLORIAM, D. E. & BABU, M. M. 2018. Pharmacogenomics of GPCR drug targets. *Cell,* 172**,** 41-54. e19.

HEDTKE, S. M., DOYLE, S. R., KUESEL, A. C., ARMOO, S., OSEI-ATWENEBOANA, M. Y., KUESEL, A. C. & GRANT, W. 2017. Multiple paths towards loss of drug sensitivity: Whole-Genome Sequencing of *Onchocerca volvulus* indicates genes under selection are dependent on transmission zonE. *American Journal of Tropical Medicine and Hygiene*

HEDTKE, S. M., KUESEL, A. C., CRAWFORD, K. E., GRAVES, P. M., BOUSSINESQ, M., LAU, C. L., BOAKYE, D. A. & GRANT, W. N. 2019. Genomic epidemiology in filarial nematodes: transforming the basis for elimination program decisions. *Frontiers in Genetics,* 10.

HEDTKE, S. 2020. *RE: host (human) DNA in O. volvulus DNA extractions.*

HEDTKE, S. 2020. *RE: Estimates of Ne based on the nuclear data from Ghana.*

HERMISSON, J. & PENNINGS, P. S. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics,* 169**,** 2335-2352.

HILL, W. & ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. *Theoretical and applied genetics,* 38**,** 226-231.

HILL, W. G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genetics Research,* 38**,** 209-216.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. & MANOLIO, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences,* 106**,** 9362-9367.

HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics,* 6**,** 95-108.

HOEKSTRA, H. E., HIRSCHMANN, R. J., BUNDEY, R. A., INSEL, P. A. & CROSSLAND, J. P. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science,* 313**,** 101-104.

HOERAUF, A., NISSEN-PÄHLE, K., SCHMETZ, C., HENKLE-DÜHRSEN, K., BLAXTER, M. L., BÜTTNER, D. W., GALLIN, M. Y., AL-QAOUD, K. M., LUCIUS, R. & FLEISCHER, B. 1999. Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility. *The Journal of clinical investigation,* 103**,** 11-18.

HOLSINGER, K. E. & WEIR, B. S. 2009. Genetics in geographically structured populations: defining, estimating, and interpreting F ST. *Nature Reviews Genetics,* 10**,** 639-650.

HOTEZ PJ. 2007. Control of onchocerciasis - the next generation. Lancet. 2007; 369(9578):1979±80. https://doi.org/10.1016/S0140-6736(07)60923-4 PMID: 17574078.

HOWIE, B., FUCHSBERGER, C., STEPHENS, M., MARCHINI, J. & ABECASIS, G. R. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics,* 44**,** 955-959.

HUANG, Y., HICKEY, J. M., CLEVELAND, M. A. & MALTECCA, C. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genetics Selection Evolution,* 44**,** 25.

HUANG, Y.-J. & PRICHARD, R. K. 1999. Identification and stage-specific expression of two putative P-glycoprotein coding genes in *Onchocerca volvulus*. *Molecular and biochemical parasitology,* 102**,** 273-281.

HUNTER, D. J. & KRAFT, P. 2007. Drinking from the fire hose–statistical issues in genome-wide association studies. *N Engl J Med,* 357**,** 436-439.

ILLUMINA 2019. Genomic solutions for cell biology and complex disease research. An overview of recent publications featuring Illumina® technology

INNAN, H. & KIM, Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences,* 101**,** 10667-10672.

INTERNATIONAL HAPMAP 3 CONSORTIUM 2010. Integrating common and rare genetic variation in diverse human populations. *Nature,* 467**,** 52.

INTERNATIONAL HAPMAP CONSORTIUM 2003. The international HapMap project. *Nature,* 426**,** 789.

INTERNATIONAL HAPMAP CONSORTIUM 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299.

ISIK, F., HOLLAND, J. & MALTECCA, C. 2017. Imputing Missing Genotypes. *Genetic Data Analysis for Plant and Animal Breeding.* Springer.

JIA, H., GUO, Y., ZHAO, W. & WANG, K. 2014. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Scientific reports,* 4**,** 1-8.

KARAM, M., SCHULZ-KEY, H. AND REMME, J. 1987. Population dynamics of *Onchocerca volvulus* after 7 to 8 years of vector control in West Africa. *Acta Trop*. 44, 445 457

KEDDIE, E. M., HIGAZI, T. & UNNASCH, T. R. 1998. The mitochondrial genome of *Onchocerca volvulus*: sequence, structure and phylogenetic analysis. *Molecular and biochemical parasitology,* 95**,** 111-127.

KHANYILE, K. S., DZOMBA, E. F. & MUCHADEYI, F. C. 2015. Population genetic structure, linkage disequilibrium and effective population size of conserved and extensively raised village chicken populations of Southern Africa. *Frontiers in genetics,* 6**,** 13.

KIM, E. S. & KIRKPATRICK, B. 2009. Linkage disequilibrium in the North American Holstein population. *Animal genetics,* 40**,** 279-288.

KLÄGER, S., WHITWORTH, J. & DOWNHAM, M. 1996. Viability and fertility of adult *Onchocerca volvulus* after 6 years of treatment with ivermectin. *Tropical Medicine & International Health,* 1**,** 581-589.

KLASSON, L., WALKER, T., SEBAIHIA, M., SANDERS, M. J., QUAIL, M. A., LORD, A., SANDERS, S., EARL, J., O'NEILL, S. L. & THOMSON, N. 2008. Genome evolution of Wolbachia strain w Pip from the *Culex pipiens* group. *Molecular biology and evolution,* 25**,** 1877-1887.

KNIERIM, E., LUCKE, B., SCHWARZ, J. M., SCHUELKE, M. & SEELOW, D. 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PloS one,* 6.

KORKUC, P., ARENDS, D. & BROCKMANN, G. A. 2019. Finding the optimal imputation strategy for small cattle populations. *Frontiers in genetics,* 10**,** 52.

KRON, M. & ALI, M. 1993. Characterization of a variant tandem repeat from Sudanese *Onchocerca volvulus*. *Tropical Medicine and Parasitology: Official Organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ),* 44**,** 113-115.

KRUGLYAK, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature genetics,* 22**,** 139-144.

LAZDINS-HELDS, J., REMME, J. & BOAKYE, B. 2003. Focus: Onchocerciasis. *Nature Reviews Microbiology,* 1**,** 178-178.

LE JAMBRE, L. F., DOBSON, R. J., LENANE, I. J. & BARNES, E. H. 1999. Selection for anthelmintic resistance by macrocyclic lactones in *Haemonchus contortus*. *International Journal for Parasitology,* 29**,** 1101-1111.

LEWONTIN, R. & KOJIMA, K. I. 1960. The evolutionary dynamics of complex polymorphisms. *Evolution,* 14**,** 458-472.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics,* 25**,** 1754-1760.

LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-2993.

LI, Y., WILLER, C., SANNA, S. & ABECASIS, G. 2009. Genotype imputation. *Annual review of genomics and human genetics,* 10**,** 387-406.

LITTLE, M. P., BASÁÑEZ, M.-G., BREITLING, L. P., BOATIN, B. A. & ALLEY, E. S. 2004. Incidence of blindness during the onchocerciasis control programme in western Africa, 1971-2002. *Journal of Infectious Diseases,* 189**,** 1932-1941.

LONG, A. D., LYMAN, R. F., LANGLEY, C. H. & MACKAY, T. F. 1998. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics,* 149**,** 999-1017.

MA, P., BRØNDUM, R. F., ZHANG, Q., LUND, M. S. & SU, G. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *Journal of dairy science,* 96**,** 4666-4677.

MACEACHERN, S., HAYES, B., MCEWAN, J. & GODDARD, M. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high-density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. *BMC genomics,* 10**,** 181.

MACKENZIE CD. 2007. Efficacy of ivermectin against Onchocerca volvulus in Ghana. Lancet. 2007; 370 (9593):1123; author reply 4-5.

MANOLIO, T. A. 2010. Genome wide association studies and assessment of the risk of disease. *New England Journal of Medicine,* 363**,** 166-176.

MANOLIO, T. A., BROOKS, L. D. & COLLINS, F. S. 2008. A HapMap harvest of insights into the genetics of common disease. *The Journal of clinical investigation,* 118**,** 1590.

MARCHINI, J. & HOWIE, B. 2010. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics,* 11.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. & DONNELLY, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics,* 39**,** 906-913.

MATUKUMALLI, L. K., LAWLEY, C. T., SCHNABEL, R. D., TAYLOR, J. F., ALLAN, M. F., HEATON, M. P., O'CONNELL, J., MOORE, S. S., SMITH, T. P. & SONSTEGARD, T. S. 2009. Development and characterization of a high-density SNP genotyping assay for cattle. *PloS one,* 4.

MCKAY, S. D., SCHNABEL, R. D., MURDOCH, B. M., MATUKUMALLI, L. K., AERTS, J., COPPIETERS, W., CREWS, D., NETO, E. D., GILL, C. A. & GAO, C. 2007. Whole genome linkage disequilibrium maps in cattle. *BMC genetics,* 8**,** 74.

MEADOWS, J. R., CHAN, E. K. & KIJAS, J. W. 2008. Linkage disequilibrium compared between five populations of domestic sheep. *BMC genetics,* 9**,** 61.

MEREDITH, S. E., LANDO, G., GBAKIMA, A. A., ZIMMERMAN, P. A. & UNNASCH, T. R. 1991. *Onchocerca volvulus*: application of the polymerase chain reaction to identification and strain differentiation of the parasite. *Experimental parasitology,* 73**,** 335-344.

MEREDITH, S. E., UNNASCH, T. R., KARAM, M., PIESSENS, W. F. & WIRTH, D. F. 1989. Cloning and characterization of an *Onchocerca volvulus* specific DNA sequence. *Molecular and biochemical parasitology,* 36**,** 1-10.

MESSER, P. W. & PETROV, D. A. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution,* 28**,** 659-669.

MONTEIRO, T. O., NETO, S. Y., DAMOS, F. S. & LUZ, R. D. C. S. 2016. Development of a photoelectrochemical sensor for detection of TBHQ antioxidant based on LiTCNE-TiO 2 composite under visible LED light. *Journal of Electroanalytical Chemistry,* 774**,** 36-41.

MUÑOZ, M., BOZZI, R., GARCÍA-CASCO, J., NÚÑEZ, Y., RIBANI, A., FRANCI, O., GARCÍA, F., ŠKRLEP, M., SCHIAVO, G. & BOVO, S. 2019. Genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds assessed with a high-density SNP chip. *Scientific reports,* 9**,** 1-14.

MURDOCH, M., ASUZU, M., HAGAN, M., MAKUNDE, W., NGOUMOU, P., OGBUAGU, K., OKELLO, D., OZOH, G. & REMME, J. 2002. Onchocerciasis: the clinical and epidemiological burden of skin disease in Africa. *Annals of Tropical Medicine & Parasitology,* 96**,** 283-296.

NANA-DJEUNGA, H., BOURGUINAT, C., PION, S. D., KAMGNO, J., GARDON, J., NJIOKOU, F., BOUSSINESQ, M. & PRICHARD, R. K. 2012. Single nucleotide polymorphisms in β-tubulin selected in *Onchocerca volvulus* following repeated ivermectin treatment: possible indication of resistance selection. *Molecular and biochemical parasitology,* 185**,** 10-18.

NEALE, B. M. 2010. Introduction to linkage disequilibrium, the HapMap, and imputation. *Cold Spring Harbor protocols,* 2010**,** pdb. top74.

NOMA, M., NWOKE, B., NUTALL, I., TAMBALA, P., ENYONG, P., NAMSENMO, A., REMME, J., AMAZIGO, U., KALE, O. & SEKETELI, A. 2002. Rapid epidemiological mapping of onchocerciasis (REMO): its application by the African Programme for Onchocerciasis Control (APOC). *Annals of Tropical Medicine & Parasitology,* 96**,** S29-S39.

NORDBORG, M., HU, T. T., ISHINO, Y., JHAVERI, J., TOOMAJIAN, C., ZHENG, H., BAKKER, E., CALABRESE, P., GLADSTONE, J. & GOYAL, R. 2005. The pattern of polymorphism in Arabidopsis thaliana. *PLoS biology,* 3.

NOTHNAGEL, M., ELLINGHAUS, D., SCHREIBER, S., KRAWCZAK, M. & FRANKE, A. 2009. A comprehensive evaluation of SNP genotype imputation. *Human genetics,* 125**,** 163-171.

OSEI-ATWENEBOANA, M. Y., AWADZI, K., ATTAH, S. K., BOAKYE, D. A., GYAPONG, J. O. & PRICHARD, R. K. 2011. Phenotypic evidence of emerging ivermectin resistance in *Onchocerca volvulus*. *PLoS neglected tropical diseases,* 5**,** e998.

OSEI-ATWENEBOANA, M. Y., BOAKYE, D. A., AWADZI, K., GYAPONG, J. O. & PRICHARD, R. K. 2012. Genotypic analysis of β-tubulin in *Onchocerca volvulus* from communities and individuals showing poor parasitological response to ivermectin treatment. *International Journal for Parasitology: Drugs and Drug Resistance,* 2**,** 20-28.

OSEI-ATWENEBOANA, M. Y., ENG, J. K., BOAKYE, D. A., GYAPONG, J. O. & PRICHARD, R. K. 2007. Prevalence and intensity of *Onchocerca volvulus* infection and efficacy of ivermectin in endemic communities in Ghana: a two-phase epidemiological study. *The Lancet,* 369**,** 2021-2029.

PAVLIDIS, P. & ALACHIOTIS, N. 2017. A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki,* 24**,** 7.

PEARSON, T. A. & MANOLIO, T. A. 2008. How to interpret a genome-wide association study. *Jama,* 299**,** 1335-1344.

PETER, B. M., HUERTA-SANCHEZ, E. & NIELSEN, R. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS genetics,* 8.

PLAISIER, A.P., VAN OORTMARSSEN, G.J., REMME, J. AND HABBEMA, J.D.F., 1991. The reproductive lifespan of *Onchocerca volvulus* in West African savanna. *Acta tropica*, 48(4), pp.271-284.

PLAISIER, A. P., ALLEY, E. S., BOATIN, B. A., VAN OORTMARSSEN, G. J., REMME, H., DE VIAS, S. J., BONNEUX, L. & HABBEMA, J. D. F. 1995. Irreversible effects of ivermectin on adult parasites in onchocerciasis patients in the Onchocerciasis Control Programme in West Africa. *Journal of Infectious Diseases,* 172**,** 204-210.

PRIEUR, V., CLARKE, S. M., BRITO, L. F., MCEWAN, J. C., LEE, M. A., BRAUNING, R., DODDS, K. G. & AUVRAY, B. 2017. Estimation of linkage disequilibrium and effective population size in New Zealand sheep using three different methods to create genetic maps. *BMC genetics,* 18**,** 68.

PRITCHARD, J. K. & DI RIENZO, A. 2010. Adaptation–not by sweeps alone. *Nature Reviews Genetics,* 11**,** 665-667.

PULASKI, C. N., MALONE, J. B., BOURGUINAT, C., PRICHARD, R., GEARY, T., WARD, D., KLEI, T. R., GUIDRY, T., DELCAMBRE, B. & BOVA, J. 2014. Establishment of macrocyclic lactone resistant *Dirofilaria immitis* isolates in experimentally infected laboratory dogs. *Parasites & vectors,* 7**,** 494.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. & DALY, M. J. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics,* 81**,** 559-575.

QANBARI, S., PIMENTEL, E., TETENS, J., THALLER, G., LICHTNER, P., SHARIFI, A. & SIMIANER, H. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Animal genetics,* 41**,** 346-356.

R DEVELOPMENT CORE TEAM 2013. R: A language and environment for statistical computing.

REDMAN, E., WHITELAW, F., TAIT, A., BURGESS, C., BARTLEY, Y., SKUCE, P. J., JACKSON, F. & GILLEARD, J. S. 2015. The emergence of resistance to the benzimidazole antihelmintics in parasitic nematodes of livestock is characterised by multiple independent hard and soft selective sweeps. *PLoS neglected tropical diseases,* 9**,** e0003494.

REICH, D. E. & LANDER, E. S. 2001. On the allelic spectrum of human disease. *TRENDS in Genetics,* 17**,** 502-510.

REICH, D. E., CARGILL, M., BOLK, S., IRELAND, J., SABETI, P. C., RICHTER, D. J., LAVERY, T., KOUYOUMJIAN, R., FARHADIAN, S. F. & WARD, R. 2001. Linkage disequilibrium in the human genome. *Nature,* 411**,** 199-204.

REMME, J.H., AMAZIGO, U., ENGELS, D., BARRYSON, A. AND YAMEOGO, L., 2007. Efficacy of ivermectin against Onchocerca volvulus in Ghana. The Lancet, 370(9593), pp.1123-1124.

REMINGTON, D. L., THORNSBERRY, J. M., MATSUOKA, Y., WILSON, L. M., WHITT, S. R., DOEBLEY, J., KRESOVICH, S., GOODMAN, M. M. & BUCKLER, E. S. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences,* 98**,** 11479-11484.

ROZAS, J., FERRER-MATA, A., SÁNCHEZ-DELBARRIO, J. C., GUIRAO-RICO, S., LIBRADO, P., RAMOS-ONSINS, S. E. & SÁNCHEZ-GRACIA, A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular biology and evolution,* 34**,** 3299-3302.

SALLÉ, G., DOYLE, S., CORTET, J., CABARET, J., BERRIMAN, M., HOLROYD, N. & COTTON, J. 2019. The global diversity of *Haemonchus contortus* is shaped by human intervention and climate. *Nature communications,* 10**,** 1-14.

SEIXAS, S., IVANOVA, N., FERREIRA, Z., ROCHA, J. & VICTOR, B. L. 2012. Loss and gain of function in SERPINB11: an example of a gene under selection on standing variation, with implications for host-pathogen interactions. *PloS one,* 7.

SERVICE, S., DEYOUNG, J., KARAYIORGOU, M., ROOS, J. L., PRETORIOUS, H., BEDOYA, G., OSPINA, J., RUIZ-LINARES, A., MACEDO, A. & PALHA, J. A. 2006. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature genetics,* 38**,** 556-560.

SHOOP, W. 1993. Ivermectin resistance. *Parasitology Today,* 9**,** 154-159.

SCHULZ-KEY, H. AND KARAM, M. 1986. Periodic Reproduction of *Onchocerca volvulus*, *Parasitol*. Today 2, 284-286

SMITH, J. M. & HAIGH, J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research,* 23**,** 23-35.

SPENCER, C. C., SU, Z., DONNELLY, P. & MARCHINI, J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics,* 5.

STEPHAN, W., SONG, Y. S. & LANGLEY, C. H. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics,* 172**,** 2647-2663.

SUDMANT, P. H., RAUSCH, T., GARDNER, E. J., HANDSAKER, R. E., ABYZOV, A., HUDDLESTON, J., ZHANG, Y., YE, K., JUN, G. & FRITZ, M. H.-Y. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature,* 526**,** 75-81.

SUTTER, N. B., EBERLE, M. A., PARKER, H. G., PULLAR, B. J., KIRKNESS, E. F., KRUGLYAK, L. & OSTRANDER, E. A. 2004. Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome research,* 14**,** 2388-2396.

SVED, J. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical population biology,* 2**,** 125-141.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585-595.

TAN, Y.-C., MICHAEEL, A., BLUMENFELD, J., DONAHUE, S., PARKER, T., LEVINE, D. & RENNERT, H. 2012. A novel long-range PCR sequencing method for genetic analysis of the entire PKD1 gene. *The Journal of Molecular Diagnostics,* 14**,** 305-313.

TENAILLON, M. I., SAWKINS, M. C., LONG, A. D., GAUT, R. L., DOEBLEY, J. F. & GAUT, B. S. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *Proceedings of the National Academy of Sciences,* 98**,** 9161-9166.

TENESA, A., NAVARRO, P., HAYES, B. J., DUFFY, D. L., CLARKE, G. M., GODDARD, M. E. & VISSCHER, P. M. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome research,* 17**,** 520-526.

TENNESSEN, J. A., BIGHAM, A. W., O'CONNOR, T. D., FU, W., KENNY, E. E., GRAVEL, S., MCGEE, S., DO, R., LIU, X. & JUN, G. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *science,* 337**,** 64-69.

THEOPHILUS, B. D. & RAPLEY, R. 2002. *PCR mutation detection protocols*, Springer Science & Business Media.

TILLER, G. R. 2014. Potential Roles of Peroxidases in *Caenorhabditis elegans* Innate Immunity.

TISHKOFF, S. A., REED, F. A., RANCIARO, A., VOIGHT, B. F., BABBITT, C. C., SILVERMAN, J. S., POWELL, K., MORTENSEN, H. M., HIRBO, J. B. & OSMAN, M. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics,* 39**,** 31-40.

TURNER, H. C., CHURCHER, T. S., WALKER, M., OSEI-ATWENEBOANA, M. Y., PRICHARD, R. K. & BASÁÑEZ, M.-G. 2013. Uncertainty surrounding projections of the long-term impact of ivermectin treatment on human onchocerciasis. *PLoS neglected tropical diseases,* 7**,** e2169.

UNICEF & UNDP/WORLD BANK/WHO SPECIAL PROGRAMME FOR RESEARCH AND TRAINING IN TROPICAL DISEASES 1996. Community-directed treatment with ivermectin: report of a multi-country study. Geneva: World Health Organization Report No. TDR/AFT/RP/96.1.

VALENTIM, C. L., CIOLI, D., CHEVALIER, F. D., CAO, X., TAYLOR, A. B., HOLLOWAY, S. P., PICA-MATTOCCIA, L., GUIDI, A., BASSO, A. & TSAI, I. J. 2013. Genetic and molecular basis of drug resistance and species-specific drug action in schistosome parasites. *Science,* 342**,** 1385-1389.

VANRADEN, P., NULL, D., SARGOLZAEI, M., WIGGANS, G., TOOKER, M., COLE, J., SONSTEGARD, T., CONNOR, E., WINTERS, M. & VAN KAAM, J. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of dairy science,* 96**,** 668-678.

VENTURA, R. V., MILLER, S. P., DODDS, K. G., AUVRAY, B., LEE, M., BIXLEY, M., CLARKE, S. M. & MCEWAN, J. C. 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genetics Selection Evolution,* 48**,** 71.

WALL, J. D. & PRITCHARD, J. K. 2003a. Assessing the performance of the haplotype block model of linkage disequilibrium. *The American Journal of Human Genetics,* 73**,** 502-515.

WALL, J. D. & PRITCHARD, J. K. 2003b. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics,* 4**,** 587-597.

WANG, J. 2005. Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 360**,** 1395-1409.

WEEDON, M. N., LETTRE, G., FREATHY, R. M., LINDGREN, C. M., VOIGHT, B. F., PERRY, J. R., ELLIOTT, K. S., HACKETT, R., GUIDUCCI, C. & SHIELDS, B. 2007. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature genetics,* 39**,** 1245-1250.

WEIR, B. S. & COCKERHAM, C. C. 1984. Estimating F‐statistics for the analysis of population structure. *evolution,* 38**,** 1358-1370.

WEISS, K. M. & CLARK, A. G. 2002. Linkage disequilibrium and the mapping of complex human traits. *TRENDS in Genetics,* 18**,** 19-24.

WICKHAM, H. 2016. *ggplot2: elegant graphics for data analysis*, Springer.

WIGGANS, G., COOPER, T., VANRADEN, P., OLSON, K. & TOOKER, M. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *Journal of dairy science,* 95**,** 1552-1558.

WIGGANS, G., COOPER, T., VANRADEN, P., OLSON, K. & TOOKER, M. 2012. Use of the Illumina B.

WINNEN, M., PLAISIER, A., ALLEY, E., NAGELKERKE, N., VAN OORTMARSSEN, G., BOATIN, B. & HABBEMA, J. 2002. Can ivermectin mass treatments eliminate onchocerciasis in Africa? *Bulletin of the World Health Organization,* 80**,** 384-391.

WORLD HEALTH ORGANIZATION & UNICEF 2018. TDR annual report 2018: Intervention and Implementation Research.

WORLD HEALTH ORGANIZATION. 2019. *Onchocerciasis* [Online]. [Accessed 7 August 2020].

WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics,* 16**,** 97.

XU, M., MOLENTO, M., BLACKHALL, W., RIBEIRO, P., BEECH, R. & PRICHARD, R. 1998. Ivermectin resistance in nematodes may be caused by alteration of P-glycoprotein homolog. *Molecular and biochemical parasitology,* 91**,** 327-335.

ZENG, W. & DONELSON, J. E. 1992. The actin genes of *Onchocerca volvulus*. *Molecular and biochemical parasitology,* 55**,** 207-216.

ZIMMERMAN, P. A., KATHOLI, C. R., WOOTEN, M. C., LANG-UNNASCH, N. & UNNASCH, T. R. 1994. Recent evolutionary history of American *Onchocerca volvulus*, based on analysis of a tandemly repeated DNA sequence family. *Molecular Biology and Evolution,* 11**,** 384-392.

ZIMMERMAN, P. A., TOE, L. & UNNASCH, T. R. 1993. Design of Onchocerca DNA probes based upon analysis of a repeated sequence family. *Molecular and Biochemical Parasitology,* 58**,** 259-267.

Appendices

Appendix for chapter three

Table 3.1. Mean pairwise LD between SNP loci in 1 Mb window in two autosomal chromosomes.

OM1			OM4		
Distance	r^2	No of SNPs	Distance	r^2	No of SNPs
	M (SD)			M (SD)	
0 -1 Mb	0.082 (0.131)	15831	0 -1 Mb	0.095 (0.193)	8349
1 - 2 Mb	0.086 (0.135)	13712	1 - 2 Mb	0.087 (0.153)	19607
2 - 3 Mb	0.106 (0.156)	8874	2 - 3 Mb	0.086 (0.146)	23683
3 - 4 Mb	0.087 (0.135)	10467	3 - 4 Mb	0.078 (0.150)	18330
4 - 5 Mb	0.106 (0.152)	9443	4 - 5 Mb	0.084 (0.147)	11959
5 - 6 Mb	0.155 (0.196)	9810	5 - 6 Mb	0.084 (0.130)	18403
6 - 7 Mb	0.122 (0.174)	10214	6 - 7 Mb	0.079 (0.128)	16846
7 - 8 Mb	0.098 (0.148)	12571	7 - 8 Mb	0.098 (0.145)	12935
8 - 9 Mb	0.088 (0.141)	17505	8 - 9 Mb	0.101 (0.150)	12292
9 - 10 Mb	0.106 (0.157)	17304	9 - 10 Mb	0.078 (0.133)	10600
10 - 11 Mb	0.094 (0.143)	12850	10 - 11 Mb	0.085 (0.154)	8163
11 - 12 Mb	0.100 (0.156)	16542	11 - 12 Mb	0.076 (0.139)	13467
12 - 13 Mb	0.101 (0.160)	10465	12 - 13 Mb	0.076 (0.136)	14960
13 - 14 Mb	0.111 (0.136)	12957	13 - 14 Mb	0.080 (0.160)	20390
14 - 15 Mb	0.084 (0.134)	20236	14 - 15 Mb	0.081 (0.148)	13923
15 - 16 Mb	0.091 (0.133)	20288	15 - 16 Mb	0.077 (0.132)	16330
16 - 17 Mb	0.084 (0.129)	18384	16 - 17 Mb	0.117 (0.163)	1487
17 - 18 Mb	0.090 (0.143)	15615			
18 - 19 Mb	0.111 (0.204)	6782			
19 - 20 Mb	0.097 (0.184)	4884			
20 - 21 Mb	0.094 (0.151)	7946			
21 - 22 Mb	0.121 (0.168)	8617			
22 - 23 Mb	0.089 (0.137)	9978			
23 - 24 Mb	0.075 (0.129)	11553			
24 - 25 Mb	0.089 (0.147)	13578			
25 - 26 Mb	0.096 (0.154)	14584			
26 - 27 Mb	0.075 (0.127)	18953			
27 - 28 Mb	0.087 (0.140)	13488			
28 - 29 Mb	0.144 (0.235)	2580			
29 - 30 Mb	0.084 (0.157)	16391			
30 - 31 Mb	0.075 (0.010)	20898			
31 - 32 Mb	0.079 (0.128)	18825			

Appendix for chapter four

Table 4.1. Data summary of the accuracy of *beagle* imputation quality metric of the four imputation combinations.

Reference panel/ Target dataset	Mean Imputation Accuracy (mean AR2)[a]		
	Rare	**Low frequency**	**Common**
	MAF <0.01	**MAF 0.01–0.05**	**MAF >0.05**
global/W. African	0.036	0.196	0.550
global/Cameroon	0.033	0.062	0.077
W. African/W. African	0.009	0.095	0.737
Cameroon/Cameroon	0.107	0.178	0.086

[a] Each cell shows the mean AR2 between true genotypes and imputed dosages for the specified MAF band and reference panel across the three bins of imputed MAF.

The W. African reference panel imputed common variants (MAF >0.05) with higher accuracy scores compared to the global reference panel on the W. African target population.