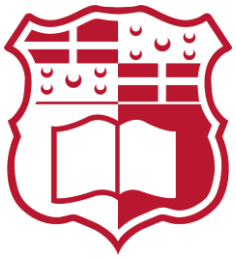

Identifying Founder Variants in the Maltese Population via Identity by Descent

Daniel Camilleri

Supervised by Prof. Rosienne Farrugia

Co-supervised by Prof. Jean-Paul Ebejer

Centre for Molecular Medicine and Biobanking

University of Malta

November 2024

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. in Bioinformatics.

GOVERNMENT OF MALTA

FONDI.eu

The research work disclosed in this publication is partially funded by the Endeavour II Scholarships Scheme. The project is co-funded by the ESF+ 2021-2027

Co-funded by the European Union

ENDEAVOUR SCHOLARSHIPS SCHEME

Acknowledgments

I would like to express my heartfelt gratitude to everyone who supported and guided me throughout this project. Without their encouragement, expertise, and patience, this work would not have been possible.

First and foremost, I am deeply grateful to my supervisors, Professor Rosienne Farrugia and Professor Jean-Paul Ebejer, for their unwavering support and insightful guidance. Your mentorship has been invaluable, from conceptualizing this project to the countless discussions that helped me navigate through many challenges. Your feedback and high standards pushed me to continually improve and to strive for excellence.

I also extend my sincere thanks to all the Masters in Bioinformatics lecturing staff, whose teaching laid the foundation for my understanding of this field. The knowledge gained from the lectures provided the essential foundation for this project and inspired me to dive deeper. Your ability to make complex topics interesting and approachable fuelled my curiosity and played a significant role in shaping the direction of my work.

My deepest gratitude goes to Professor Stephanie Bezzina Wettinger, and the MAMI Study, NGS Project and TargetID projects, for providing me with the high-throughput sequencing dataset that was essential for the computational analyses performed in this study, as well as Ms. Francesca Borg Carbott whose invaluable technical assistance throughout this project made a significant impact on its success. I would also like to thank the Endeavour II Scholarships Scheme for funding this research project.

I would like to thank my loved ones for their endless support and encouragement. Your patience and understanding allowed me the time and space to focus on this project. Your belief in my abilities gave me the strength to persevere through challenging times, and I am incredibly fortunate to have you by my side.

Each of you has played a vital role in bringing this project to completion, and I am deeply appreciative of your contributions. Thank you for helping make this journey a meaningful and fulfilling one.

Abstract

The genetic architecture of populations has been significantly shaped by various evolutionary mechanisms. Bottleneck and founder effects cause an enrichment of variants known as founder variants, and their frequency tends to be higher than expected in certain populations. Founder variants are inherited across generations as part of a haplotype block. This leads to the occurrence of a shared pattern of genetic variation between individuals who share a common ancestor, a process known as Identity-by-Descent (IBD).

The small population and geographical position of Malta made the island susceptible to many invasion and migration events, increasing the likelihood of founder effects. One Maltese founder variant has been recorded; related to 5,6,7,8-tetrahydrobiopterin deficiency (*QDPR* p.Gly23Asp). Identification of more founder variants would allow for a more specific approach in genetic testing of the Maltese population and the development of preventive measures and treatments, saving lives and resources. In this dissertation 1,076 Maltese genomes were used to analyse a list of variants and identify whether IBD segments can be used to determine the variant's founder status in the Maltese.

The benchmarking tool IBD Benchmark was used to test six IBD detection tools and identify the best performing one. This was optimised and used to perform IBD detection on a representative Maltese dataset. A list of variants of interest to the Maltese population was compiled, consisting of 15 variants locally relevant to a variety of disorders. A bioinformatics pipeline was developed to generate horizontal bar plots, heatmaps and variant genetic frameworks for founder variant analysis.

Six variants were found to be very likely founder variants of the Maltese population. The heatmaps showed these variants as part of a genetic framework, which is shared among all of the individuals that have the variant. Four variants remained inconclusive upon interpretation of the results, while no IBD segments were detected for five other variants. Having successfully met the key objectives of this study, numerous avenues for further research into founder variants have emerged. The bioinformatics pipeline developed in this project can be applied to any other variant, facilitating the process of identifying founder variants.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

AFR	African
ASMC	Ascertained Sequentially Markovian Coalescent
BH_4	5,6,7,8-tetrahydrobiopterin
CAAHD	Congenital arthrogryposis with anterior horn cell disease
cM	Centimorgan
DHPR	Dihydropteridine reductase
FSH	Follicle-stimulating hormone
GnRH	Gonadotropin-releasing hormone
GQ	Genotype quality
Hb	Haemoglobin
HMM	Hidden Markov Model
HNPCC	Hereditary nonpolyposis colorectal cancer
HPFH	Hereditary persistence of foetal haemoglobin
IBD	Identity-by-Descent
IBS	Identity-by-State
ICPP	Idiopathic central precocious puberty
IHH	Idiopathic hypogonadotropic hypogonadism
LCCS	Lethal congenital contracture syndrome 1
LD	Linkage disequilibrium
LH	Luteinizing hormone
LOD	Logarithm of the odds
MAF	Minor allele frequency
MAL	Maltese
MAMI	Maltese Acute Myocardial Infarction
Mb	Megabase
MFMD	Mediterranean founder mutation database
NFE	Non-Finnish European
NS	Nephrotic syndrome
PBWT	Position burrows wheeler transform
PCOS	Polycystic ovarian syndrome
PCR	Polymerase chain reaction
PD	Parkinson's disease
RAM	Random-access memory
ROH	Run of homozygosity
SNP	Single nucleotide polymorphism
SR	Sepiapterin reductase
STR	Short tandem repeat
TMRCA	Time to most recent common ancestor
UK	United Kingdom
U.S.	United States
VCF	Variant call file

1. Introduction

The genetics of populations has been significantly shaped by human migration. Evolutionary mechanisms, including population fusion, variation in effective population size, selection pressure, recombination rates and migration, which ultimately shape the genetic architecture, have an impact on the inheritance of variants from ancestors to modern people (Andrés and Nowick, 2014). Over the years, this has resulted in the possible selection of favourable variants over unfavourable ones (Vatsiou et al., 2016). However, variant frequencies are not only affected by these factors but also through genetic drift caused by bottlenecks and founder effects. The bottleneck effect is caused by an extreme reduction in size of a population through events such as environmental catastrophes, disease spread and famine, among others. This leads to a less diverse gene pool for that population and the frequency of the remaining variants increases (Nei et al., 1975). The founder effect, on the other hand, is caused by the geographical or cultural isolation of a few individuals from a large population. These individuals serve as the founders of a new population and are considered as the ancestors. This results in inbreeding and an overall reduction in genetic diversity, therefore causing enrichment of variants carried by the founders. Therefore, both the bottleneck and founder effects cause an enrichment of certain variants, known as founder variants, and their frequency tends to be higher than expected in certain populations (Jain et al., 2021).

As they pass through the generations, founder variants are inherited as part of a haplotype block. A haplotype block is a continuous region of the genome consisting of many single nucleotide variants that are inherited together, without any recombination events occurring within that region. This leads to the occurrence of a shared pattern of genetic variation between individuals and the segments can range from a few kilobases up to megabases in a single chromosome. This phenomenon is known as Identity-by-Descent (IBD), meaning that individuals within a population who share such regions obtained these variants through inheritance from a common ancestor. The size of the segments depends on how recent this was inherited from a common ancestor. The larger

the segment, the more recent the common ancestor. This is different to Identity-by-State (IBS), where the same genetic variant is not inherited from a common ancestor, but by chance through recombination, repeated *de novo* occurrences, or other events (Henden et al., 2018).

The founder effect can lead to an increase in the frequency of certain rare genetic variants within a population. The discovery of such variants within a population, particularly when they are pathogenic variants, allows for a more specific approach in genetic testing. One such example is the discovery of three founder *BRCA1* and *BRCA2* variants in the Ashkenazi Jewish population in the United Kingdom (UK). These variants are associated with an increased risk in breast and ovarian cancer and genetic testing is performed on a family history-based approach. The first variant that was discovered from these three is c.185delAG in *BRCA1*, which is the cause of around 16% to 20% of breast cancer cases before the age of 50 in this population (Thorlacius et al., 1997). The other two variants, the *BRCA1* c.5382insC and *BRCA2* c.6174delT affect 0.13% and 1.52% of breast cancer cases in the Ashkenazi Jewish population respectively (Roa et al., 1996). If detection of such pathogenic founder variants is performed in the entirety of the relevant population through genetic testing, many human lives and resources can be saved.

Given the small population and geographical position of Malta, the island has been susceptible to many invasion and migration events, hence increasing the likelihood of founder effects (Fiorini and Mallia-Milanes, 1991). A single Maltese founder variant has been recorded in the Mediterranean Founder Mutation Database (Charoute et al., 2015). This variant is located in the *QDPR* gene (p.Gly23Asp) and related to 5,6,7,8-tetrahydrobiopterin (BH_4) deficiency, having a carrier rate of 3.3%. The variant causes atypical hyperphenylalaninaemia and phenylketonuria, which if left untreated, can cause brain and nerve damage (Farrugia et al., 2007). Identification of such founder variants would allow for a more specific approach in genetic testing of the Maltese population and the development of preventive measures and treatments, as well as saving human lives and resources related to healthcare.

1.1 Motivation

In this dissertation 1,076 Maltese genomes will be used to analyse a list of variants and identify whether IBD segments can be used to determine their founder status in the Maltese population. Since only one founder variant has been confirmed in the population so far, one of the aims of the project will be to identify more founder variants. The compiled list of variants will contain pathogenic variants of the population, particularly those reported as having a higher frequency when compared to other populations, as well as other variants of research interest.

To achieve the first aim of this project, a comparative analysis will be performed between the most used IBD detection tools in the last five years to find the best performing one. Tang et al. (2022) developed the first open-source benchmarking method called IBD Benchmark, which can calculate the accuracy and power of these tools. In 2022, Tang et al. tested the latest developed tools of the time, which included hap-IBD (Zhou et al., 2020), iLash (Shemirani et al., 2021), RaPID (Naseri et al., 2019), TPBWT (Freyman et al., 2021) and FastSMC (Nait Saada et al., 2020). In this project, hap-IBD, RaPID, RaPID-Query, FastSMC, RefinedIBD and IBDSeq will be tested for accuracy and power using IBD Benchmark. The best performing tool will be selected to carry out IBD detection among the Maltese genomes, which will be followed by the identification of possible founder variants from the compiled list of variants. A bioinformatics pipeline will be developed to aid the process of identifying founder variants through IBD segment detection.

1.2 Aims and Objectives

The aim of this project is to perform IBD analysis, using data from 1,076 Maltese genomes, to determine whether IBD segment analysis can be used to identify the founder status of a set of variants of interest. Several tools will be used to perform this analysis. Each of these tools will be validated and compared with each other to evaluate their accuracy and performance. Furthermore, run

of homozygosity (ROH) analysis will also be performed to identify individuals that are homozygous for a haplotype block. A computational bioinformatics pipeline will also be built to interconnect the processes involved in the identification of founder variants into one programme which can be used to analyse any variant in future. This will be achieved by following these objectives:

1) Compare the performance of the IBD detection tools RaPID, RaPID-Query, FastSMC, RefinedIBD, hap-IBD and IBDSeq by recording accuracy, power and computational efficiency. The best performing tool will be optimised and used to perform IBD detection on a Maltese dataset.

2) Construct a list of variants of interest and perform IBD detection to investigate whether IBD segments can be used to classify founder variants of the Maltese population or not. ROH analysis will be performed where possible through the analysis of IBD segments in homozygous individuals.

3) Construct a bioinformatics pipeline that aids the process of identifying founder variants and ROH.

1.3 Proposed solution

Many tools have been developed throughout the years that allow for the detection of such IBD segments. A comprehensive list of these tools will be gathered, and some of the most relevant tools will be tried and tested in this project through IBD Benchmark, with the aim to use the best performing tool to identify such segments. These IBD segments will be analysed through various means, including plots, heatmaps and genetic frameworks, to find out whether a set of variants prevalent in the Maltese can be categorised as founder variants of the population. Moreover, a bioinformatics pipeline that automates the generation of the aforementioned outputs and aids the process of identifying founder variants will also be developed.

1.4 Document Structure

The subsequent “Background and Literature Review” chapter discusses founder variants in more detail, with examples from populations around the world, and their effects on public health. This chapter will also cover the different approaches that can be taken to identify such variants, based on previously documented research, particularly in relation to IBD detection analysis. A thorough description of the IBD detection tools and the benchmarking method used in this study will be given.

Chapter 3 “Methodology” will cover the methodology of the research project. This will provide a detailed description of the steps taken to achieve the aims and objectives of this project. The method of selection of the tools and datasets used will be explained, along with a description of the steps taken to develop the bioinformatics pipeline for the detection of founder variants.

In the “Results and Discussion” chapter, outputs obtained from the previously described steps will be presented and interpreted. This includes the results obtained through IBD Benchmark for the identification of the best performing tool, as well as IBD segment plots, heatmaps and genetic frameworks in relation to the variants being studied. The founder status of such variants will be determined here.

In the final chapter, “Conclusion”, the achievement of the aims and objectives of this study will be summarised. This section will also go through any limitations that were encountered, and any possible future endeavours beyond the project will be analysed.

2. Background and Literature Review

Founder variants are either caused by the geographical or cultural isolation of a few individuals from a large population, or through bottleneck effects. The latter are caused by an extreme reduction in size of a population through events such as famine, disease spread and environmental catastrophes. This leads to a less diverse gene pool for that population and the frequency of the remaining variants increases (Nei et al., 1975). When a group of a few individuals are geographically or culturally isolated from a large population, a new population is formed and these individuals serve as its founders and are considered as the ancestors. This is known as the founder effect, and it drastically reduces the gene pool and increases inbreeding within the newly formed population, causing enrichment of variants carried by the founders. Both the bottleneck and founder effect cause an enrichment of certain variants, known as founder variants, and their frequency tends to be higher than expected in certain populations (Jain et al., 2021).

Founder variants are inherited from the ancestors through generations as part of a haplotype block. This consists of a continuous region of the genome where many single nucleotide variants are inherited together, without any interference from recombination events. This leads to the occurrence of a shared pattern of genetic variation between individuals and the segments can range from a few kilobases up to megabases in a single chromosome. When a group of individuals within a population shares such regions through inheritance from a common ancestor, the phenomenon is known as IBD. The size of the segments depends on how recent this was inherited from the common ancestor, with larger segments having a more recent common ancestor. This is different to IBS, where the same genetic variant is not inherited from a common ancestor, but by chance through recombination, repeated *de novo* occurrences, or other events (Henden et al., 2018).

This chapter will cover founder variants with real world examples across different geographically and culturally isolated populations, and what factors led to the presence of such variants. Implications of founder variants on public health

and how the identification of such variants can contribute to better healthcare and economic well-being of a population will also be addressed. Moreover, the different methodologies of how founder variants are identified will be described, with emphasis on IBD detection analysis and the tools that were used in this study.

2.1 Founder Variants in Real World Populations

Several founder variants have been identified in different populations and ethnic communities around the world. Founder variants discovered within a population can be traced back to their origins and give researchers evolutionary, migration and genetic information about that population (Jain et al., 2021). Every population has a distinct genetic architecture because of its ancestry and cultural behaviour. Isolated populations, such as the Ashkenazi Jews, the Finnish (Kääriäinen et al., 2017) and the Inuit (Wallace and Bean, 2021), have been invaluable in advancing our understanding of the genetic basis of disease. Studies of founder variants in these populations have led to the identification of many disease-causing genes, providing crucial insights into genetic disorders.

2.1.1 Founder Variants in the Finnish Population

The Finnish ancestral population is built around two migration events that occurred around 4,000 and 2,000 years ago. The founding population had an estimated size of around 3,000 to 24,000, and when compared to European populations, their low genetic diversity shows that the population was isolated. The Finnish population experienced several demographic events, including rapid expansions, bottleneck effects and famine events, which shaped their genetic pool. Since then, the population experienced rapid growth, reaching an approximate figure of 5,400,000, while experiencing minimal immigration. These events led to the rise of various genetic diseases which are highly enriched in Finland. The Finnish Disease Heritage database records 43 monogenic diseases

that are more common in the Finnish than in any other population, due to the presence of founder variants (Figure 2.1) (Uusimaa et al., 2022).

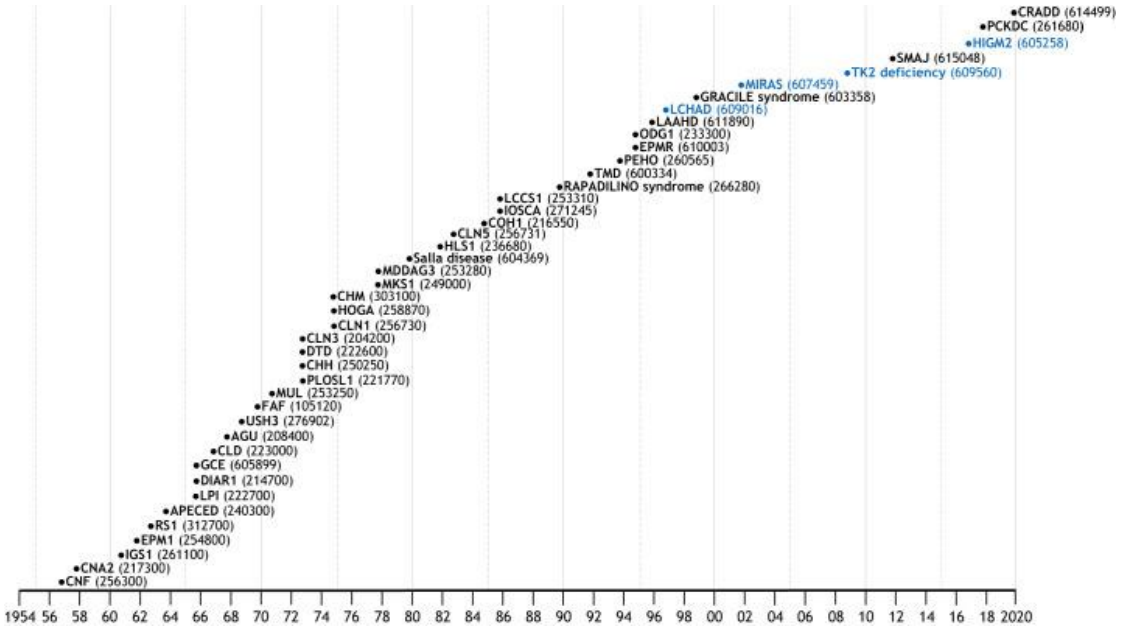

Figure 2.1: The timescale of the discovery of the genetic causes of Finnish diseases, based on their publication by the Finnish Disease Heritage. Figure reproduced from Uusimaa et al. (2022) under the terms of the Creative Commons Attribution License.

The Finnish population has been beneficial in the discovery of novel genetic loci and rare disease-causing variants which are prevalent in their population. Such examples include the discovery of *DSC1* and *SERPINB7* genes, in which atopic dermatitis causing variants have been identified (Sliz et al., 2022), and the discovery of cervical-cancer-predisposing variants in *HLA*, *CLPTM1L* and *PAX8* genes (Bowden et al., 2021). This suggests that such isolated populations with reduced genetic diversity and many founder variants provide a valuable resource for the discovery of new genetic loci and rare disease-causing variants.

2.1.2 Genetic Testing of Founder Variants in the Ashkenazi Jewish Population

Founder variants have been affecting the Ashkenazi Jewish population for over two millennia. By having a large presence in eastern and central Europe, their ancestry is derived from several different European countries. It became evident that their cultural isolation and therefore lack of genetic diversity led to

an increased prevalence of certain genetic diseases within their population, which include Fanconi anaemia type C, familial hyperinsulinism, lysosomal storage diseases, familial hypercholesterolemia, factor XI deficiency and glycogen storage disease type VII, among others (Cavalli-Sforza, 1979).

It was later discovered that breast and ovarian cancers are also prevalent in the Ashkenazi Jewish population, where they have a higher prevalence rate for three founder mutations that affect the *BRCA1* and *BRCA2* genes mainly responsible for the diseases (Rubinstein, 2004). A study on Ashkenazi Jews from Israel and the United States (U.S.) found the *BRCA1* p.E23fs variant at a carrier frequency of 0.9%, estimated to cause around 16-20% of breast cancer cases in individuals under the age of 50. The founder variants *BRCA1* p.Q1756fs and *BRCA2* p.Ser1982fs were found to have a carrier frequency of 0.13% and 1.52% respectively. Comparative testing showed that these variants are completely absent in mixed-ethnic individuals from the U.S. (Roa et al., 1996). When founder variant screening for *BRCA1/2* in the Ashkenazi Jewish population was performed, approximately 1-2.5% of women of Ashkenazi Jewish descent had one of the three founder variants (Metcalfe et al., 2010; Palomaki, 2015).

Manchanda et al. (2015) studied the cost-effectiveness of population-based testing in the UK Ashkenazi Jewish women for the three *BRCA1/2* founder variants and found that the incidence of ovarian and breast cancers could be potentially reduced by 276 and 508 cases respectively, with an overall cost reduction of £3.7 million. If the entire UK population was to undergo testing, 388 ovarian and 715 breast cancer cases have the potential to be prevented, with a cost reduction of £5.2 million. This is because medical costs for genetic testing would be far less than the costs at the later stages of a disease. This suggests that many human lives and resources can be saved if the genetic testing of these variants is extended, not only to the Ashkenazi Jewish population, but also to the entire UK population. In fact, the Dor Yeshorim genetic testing programme in the Jewish community has reduced the prevalence of certain genetic diseases in the population, including Tay Sachs disease (Ekstein and Katzenstein, 2001). Similarly, Jewish, Arab and Druze communities have also witnessed a reduction

in Thalassemia following genetic testing of 15 different beta globin variants in their communities (Koren et al., 2002). This highlights the effectiveness of ethnicity-based genetic testing programmes against highly prevalent disease-causing founder variants.

2.1.3 Founder Variants and a Case of Compound Heterozygosity in the Dutch

There are many studies across multiple European countries that have led to the discovery of founder variants. Several founder variants contributing to various diseases can be found in the younger and genetically isolated Dutch population (Kusters et al., 2011). Having experienced a bottleneck effect, 13 founder variants contributing to 13 different monogenic diseases have been identified in this population. The variants with the highest carrier frequency of 0.8% in the Dutch population are *TSEN54* p.A307S and *ABCC6* p.W1259Gfs, causing pontocerebellar hypoplasia type 2 and pseudoxanthoma elasticum respectively (Mathijssen et al., 2017). Another Dutch founder variant *MVK* p.V377I with a carrier rate of 0.65% causes HyperImmunoglobin D syndrome and is often found together with the founder variant p.I268T in the same gene. Compound heterozygosity of the two variants led to individuals suffering more often from amyloidosis (Ter Haar et al., 2016). This shows that the identification and understanding of the combined effect of founder variants may offer insights into disease mechanisms and help refine treatment strategies for them.

2.1.4 Mediterranean Founder Mutation Database

Being a central hotspot for trade routes, migration and transport since prehistoric times, the Mediterranean region is a melting point for genetic diseases. Falling at the intersection of the three continents of Europe, Africa and Asia, this region has witnessed several invasion and migration events. The originating populations of Africa first laid the genetic structure, followed by many

invasions, namely the expansion of the Phoenicians and Greeks and the Arab conquest (Capelli et al., 2006). This brought the rise of many founder variants in the region, which have been discovered across 429 studies from 20 different countries in the Mediterranean. The Mediterranean founder mutation database (MFMD) was hence created by Charoute et al. (2015) in order to compile such variants. The database currently consists of 395 founder variants, affecting 215 different genes and contributing to 224 different diseases, of which one has been reported in the Maltese. These include diseases that affect endocrine, nutritional and metabolic systems, the nervous system, the immune system, congenital malformations and chromosomal abnormalities, among others.

2.1.5 Founder Variants in the Maltese Population

Located at the centre of the Mediterranean, the Maltese population is small and geographically isolated. Having been exposed to many invasive and migration events, it is more susceptible to the presence of founder variants due to the high rate of inbreeding and reduced genetic diversity. Due to genetic similarity between the populations, the origin of the Maltese has been heavily linked to Sicily and Southern Italy, populations that primarily originated from migrating founder African populations. The small population of Malta was impacted by many invasion events from the Phoenicians, Romans, Arabs and Normans who left their genetic mark and introduced a variety of variants (Fiorini and Mallia-Milanes, 1991). At the start of the second millennium, the population was still quite small, and only recently started to grow exponentially. This led to the enrichment of many variants which were introduced by invasive populations.

So far, only one Maltese founder variant has been recorded, which can be found in the MFMD. The *QDPR* p.G23D variant which can be found in the MFMD, related to 5,6,7,8-tetrahydrobiopterin (BH_4) deficiency, was found in four unrelated Maltese patients (Farrugia et al., 2007). This molecule is involved in the homeostasis of brain neurotransmitters and hepatic phenylalanine, and variants in the Quinoid Dihydropteridine Reductase gene (*QDPR*) can result in

the deficiency of this enzyme. This causes defective recycling of BH_4, leading to atypical hyperphenylalaninaemia and phenylketonuria, causing brain and nerve damage. All the studied patients were homozygous for the causative variant *QDPR* p.G23D. The parents of the patients were all heterozygotes for the variant. Moreover, 272 random newborn DNA samples were also investigated to determine the frequency of the p.G23D variant, of which nine heterozygotes were identified. The variant has an allele frequency of 0.016 and a carrier rate of 3.3% in the Maltese, significantly higher than the global minor allele frequency (MAF) of 0.00003. Three polymorphisms were also identified, p.A32A, p.S115S and p.L132L. This establishes four distinct frameworks. The p.G23D variant was only found on framework I, the wildtype framework, and together with a high MAF value, it suggests the presence of a founder effect.

A possible founder variant having a high frequency in the Maltese is the *GNRHR* p.Q106R variant. This causes partial loss-of-function of gonadotropin-releasing hormone, which leads to luteinizing hormone (LH) and follicle-stimulating hormone (FSH) deficiency, a disorder known as idiopathic hypogonadotropic hypogonadism (IHH). These hormones are involved in the production of testosterone and oestrogen in males and females respectively and their deficiency leads to a delay or absenteeism of puberty and subsequent infertility. From 493 Maltese newborn cord-blood samples, this variant had a MAF of 0.029, approximately 10 times higher than the global population MAF of 0.003 and six times higher than the southern European population MAF of 0.005. It is suspected that this high carrier frequency is due to the variant being a founder variant of the Maltese population (Axiak et al., 2023). This variant, along with similar others, will be investigated in this project to confirm their status as founder variants.

Based on all of the previous examples, upon identification of Maltese founder variants, genetic test screening can offer many benefits if extended to such variants. Healthcare services can offer more targeted genetic practices and screening programmes to the population. People with such variants that contribute to a particular disease can get early detection and diagnosis, hence

receiving better treatment and prognosis. Moreover, this would also have a positive impact on a financial basis as medical costs for genetic testing would be far less than the costs at the later stages of a disease.

2.2 Detection of Founder Variants

Several methods are available for the identification of founder variants. These can involve either a genetics-based approach or an IBD and IBS genome-based approach. Both approaches involve the use of linkage disequilibrium (LD). By definition, LD is the non-random association of alleles at two or more loci on the same chromosome (Slatkin, 2008). In a population with random mating and no natural selection, no mutation and no migration (achieving Hardy-Weinberg equilibrium), the alleles have a random association with each other and are therefore in linkage equilibrium. If a group of alleles that are located close together on a chromosome are more frequently or less frequently found together, then they are in association and hence, in LD. In reality, the Hardy-Weinberg equilibrium is never achieved and only acts as an ideal state. Therefore, alleles will always be in LD (Ramakrishnan, 2013).

LD (D_{AB}) between alleles A and B at two loci can be calculated using the equation $D_{AB} = p_{AB} - p_A p_B$, where p_{AB} is the frequency of the AB haplotype, and p_A and p_B are the frequencies of the alleles. D has a range of value between -1 and 1, where a value of 0 indicates that the two loci are in linkage equilibrium (independent) and a value closer to -1 or 1 indicates that the two loci are in LD. If the value of D is positive, then there is a higher chance for the alleles to be found together, whereas a negative value indicates the opposite (Slatkin, 2008).

2.2.1 Genetics Based Approaches

There are three different genetics-based approaches that can be taken. The first is the single-marker analysis where LD mapping is used to compare a single genetic variant to a particular trait or disease. This method examines

whether there is a significant difference in the frequency of the allele or genotype of the variant between individuals with and without the trait or disease, and takes into account other factors such as genetic distance, recombination rates, number of generations and geographical and historical sources (Chapelle, 1993). Hästbacka et al. (1992) used this method to find a strong LD between the *DTD* gene and the CSF1R marker which is associated with diastrophic dysplasia in the Finnish population.

Similarly, the second approach multiple-marker analysis uses the likelihood estimate of gene location with multiple marker allele frequencies. This is calculated using the Malecot model. Originally, the model was created to describe kinship between two populations, however here it is modified to describe the distance between marker and disease loci (Collins and Morton, 1998). Terwilliger (1995) and Xiong and Guo (1997) found that this method can offer better resolution mapping of founder variants than the single marker analysis. In fact, Morral et al. (1994) used multiple-marker analysis to detect the founder variant *CFTR* ΔF508 related to cystic fibrosis and was able to estimate its age more accurately than with single-marker analysis.

Short tandem repeats (STRs) are one of the most common molecular markers used in genetic studies. These consist of short repeats of two or more nucleotides that form a repetitive unit (Fan and Chu, 2007). The analysis starts with the selection of chromosomal specific STR markers from a database such as STRBase (Ruitberg et al., 2001), which would be flanking the variant of interest. With the use of fluorescently labelled primers, amplification is done using polymerase chain reaction (PCR) and once amplified, the STR fragments are separated by size using capillary electrophoresis. The size of such fragments can be measured, corresponding to the number of repeat units. The number of repeats remains consistent in cases that have the same founder variant. In contrast, the repeat numbers vary in negative control cases that do not have the variant. If a STR marker is always found to be associated with the adjacent variant, this contributes towards confirming the founder status of the variant (Almeida and Korch, 2004). Mejri et al. (2012) conducted a STR marker analysis

in relation to β-thalassaemia major and β-thalassaemia intermedia, and found that the STR marker D14S72 is specific to the former disease form while D14S990 and D14S68 are related to the latter.

The third approach, ancestral haplotyping, involves the use of several analyses, such as genealogical studies and linkage and association analyses. An association is performed between a trait or disease of interest and a genomic marker, by studying the haplotypes and genomic history of a group of affected individuals from a population, of which their families are known to have a history for this particular trait or disease. Clinical and ancestral data of these individuals is required. Genotyping is then performed and linkage and allele association studies are performed on a set of markers, suspected of causing the trait or disease of interest. Nyström-Lahti et al. (1994) took this approach to study 18 Finnish families with a familial history of hereditary nonpolyposis colorectal cancer (HNPCC), which is responsible for up to 13% of colorectal cancers. This resulted in the identification of a 10 centimorgan (cM) haplotype around the *HNPCC* locus in five of the nine families on which linkage studies were possible. The cM is a unit of genetic distance between two positions in the same chromosome. If the chance of recombination between two loci is 1%, the loci are said to be 1cM apart. On average, 1cM is equal to about 1,000,000 basepairs (1 megabase, Mb), but recombination rates vary throughout the genome, so in low-recombination regions 1cM may span up to 5Mb. These five families were also found to have a shared ancestry. The same haplotype was also found on two other families in which linkage analysis was not possible. This is suggestive of an ancestral founding mutation around the *HNPCC* locus. Virtaneva et al. (1996) also took such approach in a study consisting of 53 unrelated Finnish families and found an association between the *EPM1* gene responsible for progressive myoclonus epilepsy, and the haplotype markers D21S2040 and D21S1259 in the founder Finnish population.

2.2.2 IBD and IBS Genome-based Approach

The genome-based approach is a more modern approach where pathogenic genetic founder variants are analysed in unrelated patient genomes. These can be compared to examine whether they share the same haplotype, which depends on the LD of the variants that flank the founder variant. This is known as IBD and IBS analysis. Alleles in LD which tend to be inherited together more often than expected and with minimal crossing over during meiosis start to form haplotype blocks. In IBD, unrelated individuals have the same haplotype block that descends from a common ancestor, hence carrying the founder variant. In IBS, individuals can also have the same haplotype block and variant, but do not share a common ancestor. This can happen by chance due to random shuffling of genetic material or *de novo* occurrences of the same variant in different individuals (Henden et al., 2018). If a segment is IBD, it originates from the ancestors of the population and will be found in many population individuals. Since IBS segments arise by chance, there will not be many occurrences of the segment within the population. IBD and IBS segments can be detected through several IBD detection tools that are publicly available (Jain et al., 2021).

As with founder effect analysis, IBD detection also has other applications, such as the prediction of genotype frequencies, estimating genetic variance, gene mapping and predicting inbreeding depression (Sticca et al., 2021). By knowing the length of the IBD segment, one can also estimate the age of the variants and the most recent common ancestor. The longer the IBD segment, the more recent the common ancestor is. Tools such as DMLE+ (Reeve and Rannala, 2002) are able to infer such calculation. If the boundaries of the IBD segment are known and demographic information about the population is available (growth rate, population size and MAF of the variant), DMLE+ is able to infer the age of the segment in generations. This in turn can be used to confirm any founder variants and estimate a population's demographic history over time, including population size, bottlenecks and subsequent founder effects (Sticca et al., 2021).

2.3 IBD Detection Tools

Table 2.2 provides some information about the most commonly used IBD detection tools in literature from the last five years (2019-2024). These tools use several different algorithms to detect IBD segments, some of which include seed-and-extend algorithms, Hidden Markov Models (HMM), Position Burrows Wheeler Transform (PBWT) and hashing.

Table 2.1: The most used IBD detection tools in literature from the last five years.

Tool	Brief Description	Reference
IBDSeq	Uses a probabilistic algorithm method to estimate IBD segments by using a logarithm of the odds (LOD) score.	Browning and Browning (2013)
hap-IBD	Uses position burrows wheeler transform (PBWT), seed-and-extend algorithms and allows multi-threaded IBD analysis to increase computational efficiency.	Zhou et al. (2020)
GERMLINE	Identifies identical genomic regions and extends them to determine the IBD segment.	Gusev et al. (2009)
FastSMC	Uses an improved and more efficient GERMLINE algorithm, also known as GERMLINE2, to detect IBD.	Nait Saada et al. (2020)
FastIBD	Finds short IBD segments. Superseded by Refined IBD.	Browning and Browning (2011)
Refined IBD	Uses a combined FastIBD and GERMLINE algorithm to find the extended IBD segment. Also uses a LOD score for the identified segments.	Browning and Browning (2013)
Ilash	Uses local sensitive hashing to identify small DNA segments and uses hash values to compare and identify pairs of individuals who possibly share IBD.	Shemirani et al. (2021)
RaPID	Uses PBWT and considers exact sequence match as IBD.	Naseri et al. (2019)
TRUFFLE	Identifies IBD segments and calculates the average IBD shared between individuals.	Dimitromanolakis et al. (2019)
IBIS	Uses relatedness inference to infer IBD segments.	Seidman et al. (2020)
IBLDL	Uses an HMM with a background LD model to estimate IBD segments.	Han and Abney (2011)
RELATE	Uses a continuous time Markov model to identify IBD segments.	Albrechtsen et al. (2009)

The six most commonly used IBD detection tools in literature include hap-IBD (Zhou et al., 2020), RefinedIBD (Browning and Browning, 2013), FastSMC, (Nait Saada et al., 2020), RaPID (Naseri et al., 2019), RaPID-Query (Wei et al., 2023) and IBDSeq (Browning and Browning, 2013). Therefore, these tools were

tested to find the best performing one and ultimately perform IBD detection on a Maltese dataset to determine the founder status of a list of highly prevalent variants.

Apart from the tools in Table 2.2, there are several other IBD detection tools that have not been mentioned in literature for IBD detection in the last five years. These include SILO (Wang et al., 2023), HapFABIA (Hochreiter, 2013), Parente (Rodriguez et al., 2015), FISHR (Bjelland et al., 2017), diCal-IBD (Tataru et al., 2014), KING (Manichaikul et al., 2010), TBPWT (Freyman et al., 2021), IBD_Haplo (Brown et al., 2012) and HaploScore (Durand et al., 2014).

2.3.1 hap-IBD

The Java implemented tool hap-IBD uses a PBWT (Durbin, 2014) algorithm and a seed-and-extend technique to identify IBD segments. Just like the standard Burrows-Wheeler Transform, PBWT creates a matrix of cyclic rotations by first permuting the input sequence, and then sorts these rotations in lexicographic order and extracts the last column of the sorted matrix, which forms the transformed sequence. This groups similar sequences together, hence promoting compressibility of the data. Additionally, the positional information of this data is also maintained. This provides an efficient way to store and retrieve such data. PBWT also allows for parallelization, hence also making it more time efficient.

hap-IBD does not require any pre-processing steps and takes variant call files (VCFs) by default. A genetic map file in Plink (Purcell et al., 2007) format is required. Once the VCF is fed to the tool, the PBWT algorithm goes through each marker in chronological chromosome order, and for every marker, the reverse haplotype prefixes are sorted lexicographically. This efficiently identifies all seed IBD segments. The tool tries to extend the segment if another long seed for the same haplotype pair is separated by a short IBS gap. By default, the tool allows a minimum value of 1cM and 1,000 basepairs in length for these extensions, and these values can be changed by the user. hap-IBD tries to extend every seed if

possible, and the same segment can be extended more than once until it can no longer be extended. This makes the tool more robust against possible gene conversions and genotype errors. The tool has a minimum output length of 2cM by default, which can be changed by the user. Other optional parameters are also offered by the tool, including a min-markers parameter where the identified segments need to have a minimum number of markers. hap-IBD provides two outputs, one containing between-individual IBD segments and another containing within-individual ROH segments (Zhou et al., 2020).

2.3.2 RefinedIBD

Like hap-IBD, RefinedIBD is a Java implemented tool. This tool uses the GERMLINE (Gusev et al., 2009) algorithm, together with a probabilistic approach to detect IBD segments. The tool also employs Beagle's (Browning and Browning, 2007) haplotype phasing method, where the haplotypes of the inputted VCF are phased and missing data is inputted. The first step of IBD detection applies GERMLINE's dictionary approach, which does not allow any mismatch of alleles between shared haplotypes. The algorithm identifies shared segments between pairs of individuals by detecting consecutive markers that match above a certain threshold specified by the user. In the second step, Beagle's Hidden Markov Model is used.

The phased haplotypes are used to build a haplotype frequency model, from which candidate IBD segments are identified. Then, the likelihood logarithm of the odds (LOD) score, which is the base 10 log of the likelihood ratio, is calculated for each candidate segment. By default, candidate segments having a score greater than the specified threshold, or 3.0 by default, are counted as IBD segments and given as output. The tool also uses a default minimum cM length of 1cM for the segments, unless changed by the user. RefinedIBD also allows the option to include a genetic map file in Plink format for better marker mapping and outputs cases of ROH as well in a separate output file (Browning and Browning, 2013).

2.3.3 FastSMC

FastSMC uses a hashing algorithm to identify IBD segments, together with a coalescent-based HMM to verify such segments. Implemented using C++ and optional Python bindings, the tool requires for the VCF to be converted into an Oxford phased haplotype file (.hap/.hap.gz, .samples). This can be done by using the *convert* command paired with the *--hapsample* subcommand found in BCFtools (Danecek et al., 2021). A genetic map needs to be prepared to have the exact sites as in the original VCF. The identification step of FastSMC is adapted from the GERMLINE algorithm and developed into GERMLINE2. For a given window (*w*), the algorithm divides the inputted haplotype data into windows of either 16 or 32 single nucleotide polymorphisms (SNPs), depending on the available memory resources. Every haplotype is converted into binary sequences and effectively hashed into groups of identical segments. This step is repeated for *w* + 1 for each bin that has more individuals than a predetermined threshold (i.e. a low complexity bin) until no more low complexity bins are identified. Then, every pair of individuals sharing a bin is noted and inputted in a different hash table that contains potential segments. Pairs of individuals that share sufficiently lengthy contiguous windows are reported and submitted for validation (Nait Saada et al., 2020).

After the identification step, the identified segments are fed into the Ascertained Sequentially Markovian Coalescent (ASMC) algorithm to be validated. This coalescent based HMM model estimates the time to most recent common ancestor (TMRCA) for a pair of individuals at each marker using sequencing or array platforms. The algorithm takes demographic information from a decoding quantities file which needs to be inputted by the user, using it to improve the accuracy in detecting regions of low TMRCA. The HMM emits probabilities that correspond to the probabilities of finding both genotypes of the analysed pair of individuals and variant frequencies given the TMRCA at each site. According to the Simonsen-Churchill model, transitions between hidden states in the HMM correlate to changes in TMRCA along the genome that are

caused by recombination events. The inputted decoding quantities file is then used to obtain the demographic history of the analysed haplotypes and to calculate state distributions and transition and emission probabilities. Using dynamic programming, the most likely posterior sequence of TMRCAs in the genome inferred. For each candidate segment, the posterior probability of the coalescence time is calculated. This represents the time when two lineages in genealogy merge into a common ancestor. The algorithm checks if the posterior probability of the coalescent time between the present time and the user-specified time is higher than the prior position. If the posterior probability at a site satisfies the threshold condition, the site is considered to be part of an IBD segment. The algorithm then extends the IBD segment to the next site until the condition is no longer satisfied, at which the segment breaks. The higher the average probability, the more likely the segment is IBD. As an output, FastSMC includes the age estimate and the IBD quality score of every reported segment. The tool contains some optional parameters which can be changed by the user, such as IBD segment length and various output style options (Nait Saada et al., 2020).

2.3.4 RaPID

Similar to hap-IBD, RaPID (Naseri et al., 2019) also uses the PBWT algorithm. However, since PBWT does not allow exact or single variant mismatches, low resolution random projections of the original sequences are first produced by RaPID and combined with PBWT's results. This provides an added benefit which accounts for some marker density and error rates, based on the parameter configurations given by the user. These include *r*, *s* and *w*, which stand for the number of runs, the minimum number of runs required for a match to be taken into consideration and the number of SNPs per window respectively. In the first step of RaPID, the input genotype panel is subjected to multiple random projections. The panel is divided into windows based on *w*, and within each window RaPID selects a variant site based on the highest minor allele frequency. This is followed by PBWT which identifies exact matches above a certain user

specified threshold. A hit is counted if it occurs at least *s* number of times per *r* number of runs. Since the start and end positions of two hits from two different runs will unlikely be the same, the overlapping segment of the hits is taken. The tool requires a genetic map file, which can be prepared with the use of two python scripts that are provided with the tool. The first script *filter_mapping_file.py* filters the inputted Plink genetic map file, followed by the interpolation of loci when using the second script *interpolate_loci.py*, based on the respective VCF used. The output consists of a file containing reported IBD segments. The minimum length of detected IBD segments (*d*) is also mandatory and needs to be specified by the user.

2.3.5 RaPID-Query

RaPID-Query (Wei et al., 2023) uses a combination of two existing methods, RaPID's algorithm of detecting IBD segments and a modified version of the PBWT algorithm, different than the one used in RaPID. Sanaullah et al.'s (2021) version of the original PBWT algorithm introduced by Durbin (2014) uses a sweep match query algorithm where the query haplotype is inserted into the panel. The location of the query haplotype is tracked along the panel until a match block is found. A match block is created if the cutoff long match length *L* is at least equal to the length of the start position e of the set-maximal match to the current scanned site location *k*. The block is then extended in both directions, based on the divergence values of the neighbouring haplotypes. The block is extended only if the divergence value indicates that the length of the neighbouring haplotype outside of the match block edge and the haplotype in the match block edge is at least $k + 1 - L$.

The newly modified x-PBWT-Query method eliminates the tracking of the query haplotype, as well as the divergence values if they are already present in the set-maximal match block. When the constraint $k + 1 - L$ is met, the match block includes all haplotypes in the set-maximal match block. The algorithm also uses a site distance tracking feature, which enables it to handle long match length

cutoffs in either physical or genetic units. This feature tracks the site index *i* that is *L* units away from the currently scanned site *k*, while maintaining the same algorithmic complexity. It keeps track of the closest site *i* that is *L* cM distance away from site *k*, updating this information as the site scanning progresses. Since the scanning process only moves forward, the maximum number of updates is bounded by the number of sites, resulting in a linear operation. When expanding a match block, the site distance track index *i* marks the start position of the matches within the block. In the x-PBWT-Query algorithm genetic maps are applied to the distance tracking variable *i* to facilitate the querying of the panel using the genetic distance cutoff length *L* (Wei et al., 2023).

Since the PBWT algorithm does not allow for any mismatches, which can arise due to genotyping error, mutation or gene conversion, RaPID-Query uses an algorithm similar to RaPID's. The same parameters of the number of runs (*r*), the minimum number of runs required for a match to be taken into consideration (*c)* and the number of SNPs per window (*w*) are used. The haplotype sequences are divided into multiple lower-resolution sequences of *w* equal sizes, where for each window a variant site is sampled according to the highest allele frequency. This needs to match the minimum IBD markers in number of sites (*lm*) parameter specified by the user. Exact matching segments are found from these low-resolution panels and compared to the full-resolution panels, the latter chosen by the minimum IBD markers in number of sites for high resolution (*lmh*) and the minimum IBD length in cM for high resolution (*dh*) parameters. The full-resolution panels are used to refine the boundaries of the low-resolution segments, and IBDs are merged if the distance between them is within the allowed distance *dg*. An IBD is counted if it occurs at least *c* number of times per *r* number of runs. The minimum length of detected IBD segments (*d*) also needs to be specified (Wei et al., 2023).

2.3.6 IBDSeq

IBDSeq (Browning and Browning, 2013) is a Java based tool which uses a probabilistic algorithm to estimate the observed genotypes with error for each pair of individuals. The first step of the algorithm involves variant filtering of the VCF, where all of the variants with one minor allele carrier are eliminated. Then, for each variant, the squared-correlation for the per-sample minor allele count is calculated between the said variant and each of the 250 previous variants. If this value exceeds a specified threshold (r^2 = 0.15 by default), and if none of the variants have been previously marked as excluded, the variant is marked with the higher MAF as excluded. This is followed by the computation of single-marker LOD scores. By default, segments with a LOD score of 3 or higher are counted as IBD, unless stated by the user. IBDSeq requires the use of unphased genotypes, therefore unphasing of the VCF needs to be performed (Appendix A). Optional parameters include the aforementioned minimum LOD score and r^2 value, as well as some output style options. Unlike the other tools, IBDSeq does not calculate the cM distances of the IBD segments.

2.4 IBD Benchmark

Over the years, no independent review of IBD detection tools has been performed to determine the best performing one, until the release of IBD Benchmark by Tang et al. (2022). IBD Benchmark is an open-source tool that can be used to compare IBD detection tools by calculating multiple measurements related to accuracy and power, using direct and reproducible methods. These accuracy and power metrics defined in this tool consider both coverage and length in single and multiple IBD segments. Tang et. al tested RaPID (Naseri et al., 2019), FastSMC (Nait Saada et al., 2020), TPBWT (Freyman et al., 2021), hap-IBD (Zhou et al., 2020) and iLash (Shemirani et al., 2021), and concluded that all of the tools had high accuracy for long segments (>5cM). For shorter segments, hap-IBD, iLASH and FastSMC had the highest accuracy, but were affected with the introduction of genotyping errors.

2.4.1 Simulated Datasets

To evaluate the performance of the tool, ground-truth IBD segments are required. These are the true segments of the test dataset, and the results of any IBD detection tool are compared to them to evaluate the tool's performance. Ground-truths in real data are hard to come by and are very limited, mostly to close relatives and long IBD segments. Moreover, the genealogy of such individuals is required to confirm them. However, with genetic coalescent simulators one can simulate a dataset based on real-world populations and generate its ground-truth segments. This also promotes direct comparability and reproducibility in the results. Tang et al. (2022) used the coalescent simulator tool msprime v.1.0.1 (Kelleher et al., 2016) to simulate a number of IBD datasets and generate their ground-truths. msprime simulates the ancestry and genetic diversity of individuals backwards in time by using certain population factors such as population size changes, migration and bottleneck effects. It also considers LD, segregation and mutations, and generates a whole population tree for them. Its memory-efficient algorithm also makes the tool computationally efficient to generate large and complex datasets.

Tang et al. (2022) simulated four population datasets: European , African, East Asian and a mixture of the three. These consisted of chromosome 20 sequences of 4,000 individuals each and was based on the out-of-Africa population model created by Gutenkunst et al. (2009), which contains information about the human expansion out of Africa and how the new world populations developed from it. It takes into account several migration and bottleneck events and tries to predict the non-synonymous variation across the new populations. HapMap's phase II GRCh37 map (Frazer et al., 2007) was used as a recombination map and a standard mutation rate of 1.38×10^{-8} was used to generate the datasets. Ground-truth IBD segments were generated using msprime. From the generated sequences, trees were sampled for every 5,000 basepair distance and true IBD segments were only extracted if their genetic

lengths were equal to or larger than 1cM. After dataset simulation, sites containing singletons and multiple allele values were filtered out.

Array density datasets were also generated by down sampling the original sequencing datasets. These consisted of 17,197 markers, the same number of markers on chromosome 20 of the UK Biobank data (Sudlow et al., 2015). A cM interval (*I*) that indicates the ideal distance between two markers was calculated by dividing the total chromosome genetic length (in cM) by the number of sites. This however required a window size as there may not be any markers in continuous cM intervals, and this was calculated by taking *w* markers from a *w*-by-*I* range. As a result, IBD Benchmark first tries to find a marker from its original cM interval, and if none are found, it takes a marker from the *w*-by-*I* leftover range. When selecting a marker, the program takes the one with the highest MAF, which is defined as the most frequent allele. Tang et al. (2022) tested multiple window sizes, with the window size of 5 giving the best results.

2.4.2 Genotype and Phasing Error Simulation

When the genotype of an individual observed through molecular analysis differs from the true genotype, this is known as genotyping error. These errors can be generated through every stage of the genotyping process, starting from sampling and DNA extraction to molecular and data analysis, as well as chance and human error. This can cause a reduction in detection power in downstream analysis, and needs to be accounted for (Bonin et al., 2004). To simulate them, Tang et al. (2022) generated copies of the simulated datasets with randomly introduced genotyping errors. The rates included 0.1%, 0.2%, 0.3% and 0.4% for both sequencing and array data. Additional genotype error rates of 0.0125%, 0.025% and 0.05% were also included for sequencing data as this represents better the high accuracy of the dataset when compared to array datasets.

Similarly, phasing errors occur when variants are assigned to the incorrect haplotype. Although to a much lesser extent, phasing errors can also reduce the power of downstream analysis. Therefore, phasing errors were also simulated in

the datasets to investigate their effect on the IBD detection tools. The HapMap genetic map (GRCh37) (Frazer et al., 2007) of chromosome 20 was used to simulate a population of 2,000 Europeans using stdpopsim (Adrion et al., 2020) and its OutofAfrica_2T12 model, and 0.1% genotyping errors were simulated in the population. SHAPEIT4 (Delaneau et al., 2019) was used to phase the genotype data. Tang et al. (2022) calculated an average phasing error rate (switching error) of 0.17% using VCFtools (Danecek et al., 2011).

2.4.3 Accuracy and Power Metrics

The tool's ability to detect IBD segments is evaluated using seven measures. The first measure, accuracy, is calculated by dividing the number of covered reported IBD segments by the total number of reported segments. By default, IBD Benchmark uses a 50% cutoff to calculate these measures, therefore a covered IBD segment is counted if a ground truth segment covers at least 50% of its length. The second measure, length accuracy, is calculated by taking the best-matching ground-truth segment which covers the reported segment with the highest overlap, and calculating the percentage covered. An average percentage coverage is then taken across all reported segments. This accounts for any false positive reported segments. The third measure, length discrepancy, is calculated by the root-mean-square deviation in length of the reported IBD segment and the best-matching ground-truth IBD segment. This measure is calculated in cM and the quality increases the smaller the deviation in length (Tang et al., 2022).

Recall is calculated by the proportion of the number of ground-truth IBD segments that have been reported by the tool. If the reported segment covers at least 50% of the ground-truth segment, as per the cut-off, then it is assumed to be detected. The power measure is calculated by the average proportion of ground truth IBD segments covered by the best-matching reported segment, the latter being the segment that has the most overlap. The measure of accumulative power is similarly calculated as the power measure, but any reported IBD

segments that overlap with the ground-truth segment are taken into account at the same time. Similarly, the accumulative recall measure is calculated as the recall measure, but takes into account all the reported segments that cover at least 50% of the ground truth segment (Tang et al., 2022).

Figure 2.2 demonstrates how IBD Benchmark calculates the accuracy and power of IBD detection tools. The blue segments represent the ground truth and reported IBD segments, the green segments represent the considered segments for the calculations and the grey segments are not considered in the calculations. The tool calculates these metrics across different cM bins as follows: (2, 3), (3, 4), (4, 5), (5, 6), and (7, ∞) cM bins. This makes it possible to know the performance of the tool at different cM lengths of IBD segments (Tang et al., 2022).

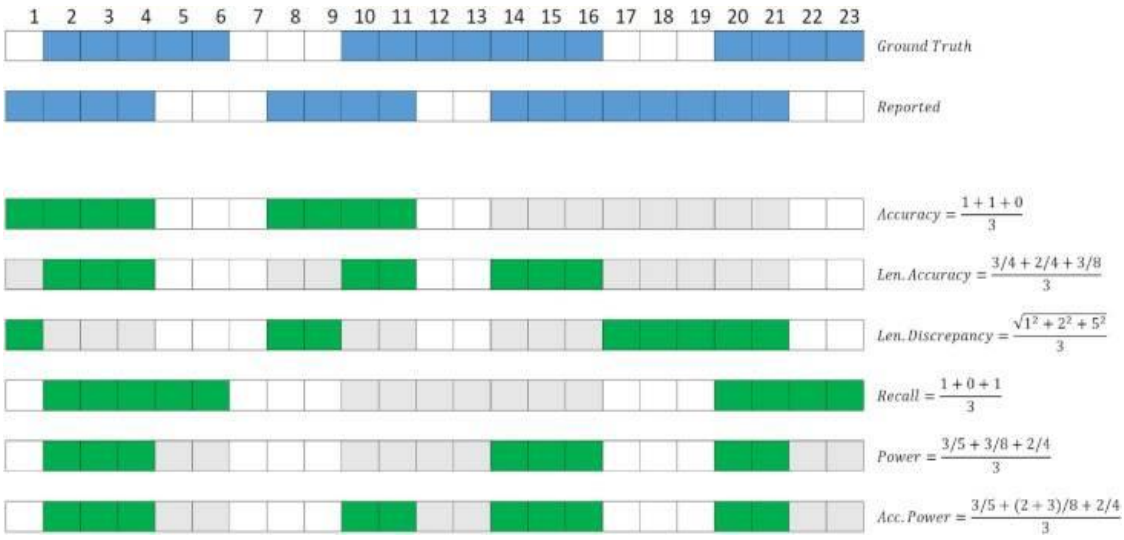

Figure 2.2: Accuracy and power metrics. The reported and ground-truth IBD segments are represented at the top in blue. The segments that are considered in calculations are marked in green, whereas those marked in grey are not considered. Figure reproduced from Tang et al. (2022) under the terms of the Creative Commons Attribution License.

2.4.4 Benchmark Results

Tang et al. (2022) tested out five of the most recent IBD detection tools on the generated sequencing and array datasets using the IBD Benchmark tool. These include RaPID (Naseri et al., 2019), FastSMC (Nait Saada et al., 2020), TPBWT (Freyman et al., 2021), hap-IBD (Zhou et al., 2020) and iLash (Shemirani et al., 2021). The results concluded that all the tools had high accuracy for long

segments (>5cM). For shorter segments, hap-IBD, iLASH and FastSMC had the highest accuracy, but were affected with the introduction of genotyping errors. RaPID was the most robust tool, maintaining high accuracy and power across the various genotype error rates.

Phasing errors did not have a significant effect on the overall power and accuracy of the tools. For shorter segments ranging from 2 to 3cM, the reduction in power was from 4% to 7% among the tools. This was greater for very long segments (≥15cM), with a reduction in power ranging from 5% to 20%. However, larger segments are less commonly found and should not have a significant effect on downstream analysis. A minor reduction in accuracy and length accuracy was also noted among the tools. This comes down to the high accuracy of current phasing algorithms and the robustness of IBD detection tools against phasing errors (Tang et al., 2022).

2.4.5 Run Time and Memory Consumption

Tang et al. (2022) tested the computational efficiency of the tools using UK Biobank's chromosome 1, consisting of 487,409 individuals and 53,260 markers, and a similar simulated dataset of 500,000 individuals containing 51,190 markers. The total size of each file was 100 gigabytes (GB). These were tested out on a server containing 500 GB of memory, except for FastSMC as it required the latest updated operating system and libraries, which were not available. Due to limited resources, this was run on a 32 GB memory PC and smaller sample sizes ranging from 1,000 to 31,000 individuals. The estimated run time and memory usage of the tool was calculated by extrapolation using second-order polynomial regressions. According to the estimations, this would have a run time of 126 days and a memory consumption of 6.5 terabytes. TPBWT (Freyman et al., 2021) was not run on the UK Biobank data due to license conflicts, therefore its results were based on the simulated dataset.

The results showed that hap-IBD had the shortest wall clock time of 0.5 hours for IBD segment detection, followed by iLash, RaPID and TBPWT. In terms

of memory consumption RaPID requires the least amount. Its maximum memory usage was limited to less than 8 GB, followed by hap-IBD at 112 GB, iLash at 191 GB and TBPWT at 488 GB. Both TPBWT and iLash were unable to finish the processing of the largest datasets with 500 GB of memory. (Tang et al., 2022).

2.5 Summary

Given the small population and isolated geographical position of Malta, the Maltese population is susceptible to the presence of founder variants. With only one founder variant identified in the population so far, related to 5,6,7,8-tetrahydrobiopterin deficiency with a carrier rate of 3.3% (Farrugia et al., 2007), further population studies are required to identify more. Studies on isolated populations like the Finnish, Ashkenazi Jews and the young Dutch population give an insight on the benefits of founder variant analysis, including the discovery of new genetic loci and rare disease-causing variants, the effectiveness of ethnicity-based genetic testing programmes, disease mechanisms and treatment strategies.

In this chapter, a brief overview of these populations was given. The different methods of founder variant identification were described, namely the use of haplotype blocks in IBD and IBS detection and the availability of tools for this process. The application of IBD Benchmark to compare the power and accuracy of IBD detection tools was also described.

3. Methodology

This chapter will go through the design and implementation of our bioinformatics pipeline, with the aim of investigating founder variants in the Maltese population. The following is discussed in this chapter:

- A representation of the collected variants that will be explored in this project for their founder status.

- The method of assessment used to select the best performing IBD detection tool with the use of test data and IBD Benchmark analysis.
- A description of the dataset used, as well as the preprocessing steps taken to prepare the dataset for IBD detection, including filtering steps, chromosomal extraction and haplotype phasing.
- The Python script used to develop and output plots, heatmaps and genetic frameworks for any variants classified as founder variants, as well as a text file output containing the variant allele genetic framework.

3.1 Summary of the Bioinformatics Pipeline

Figure 3.1 presents an overview of the bioinformatics pipeline and steps taken to develop it. This includes the scripts and tools used to generate results.

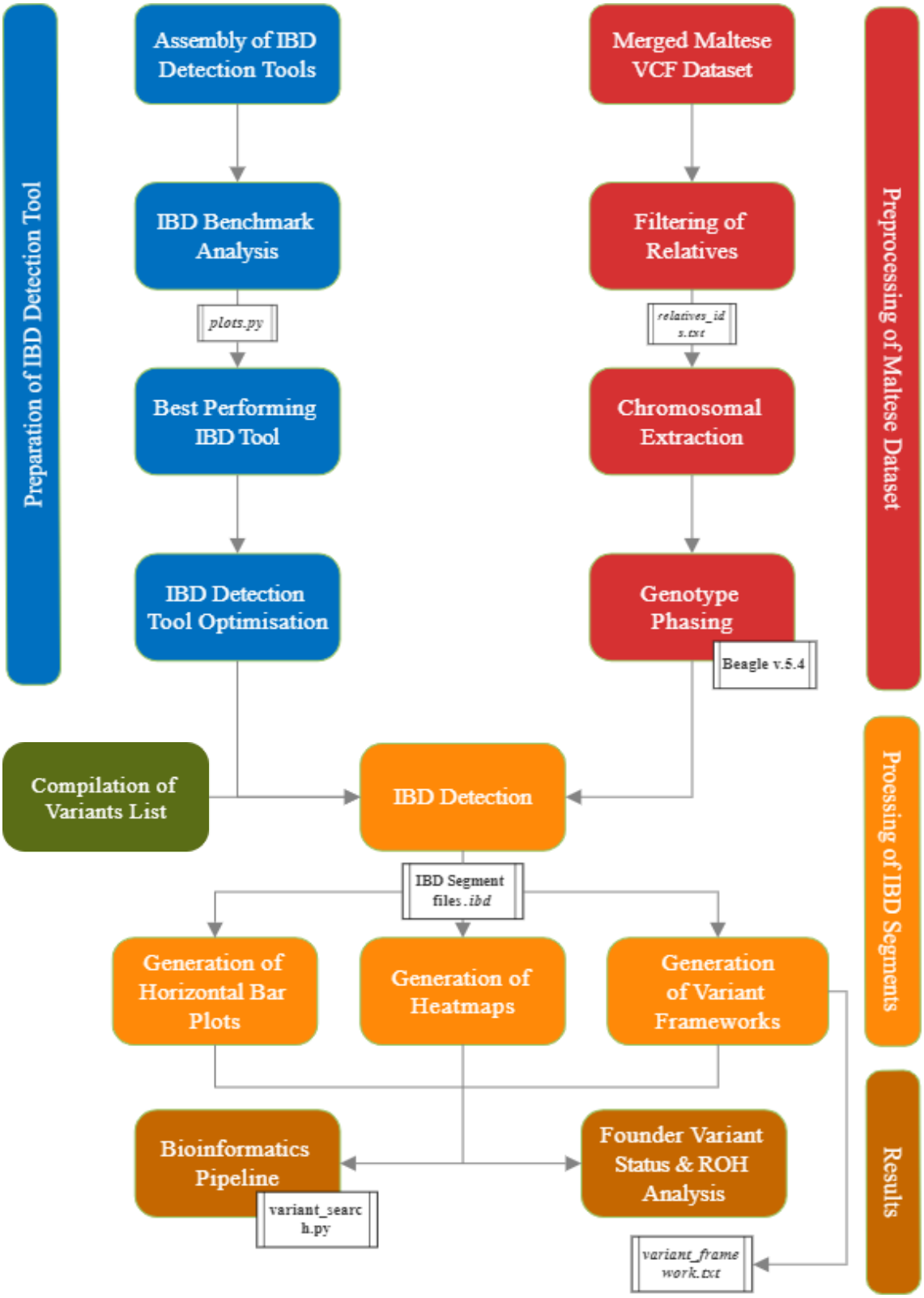

Figure 3.1: Overview of the process involving IBD tool selection, dataset handling and IBD detection for the identification of founder variants.

3.2 Technology Stack

Python 3 (Rossum and Drake, 2009) is the primary computing language that was used for bioinformatics analysis. Python has emerged as a popular choice for bioinformatics analysis due to its ability to handle large datasets and its flexibility in creating algorithms to manipulate the data, making its versatility and reliability an ideal choice for bioinformatics analysis. A bioinformatics pipeline was built using this programming language. The aim of the pipeline is to automate and interconnect the processes involved in founder variant analysis. The Python version 3.9 was used. All of the code that was developed and used throughout the project can be found in the following GitHub repository: https://github.com/danielc1999/Dissertation-Founder-Variants

Accuracy and power testing of the six IBD detection tools was performed on an Ubuntu v.20.04.6 server with 1 TB of random-access memory (RAM) and managed by the University of Malta. The accuracy and power plots were plotted using the Python script *plots.py*.

IBD detection of the Maltese dataset was done with the best performing tool on an Ubuntu v.20.04.2 server with 791 GB of RAM and managed by the University of Malta. The bash script *generate_ibds.sh* was used to filter, extract and phase the dataset, as well as identify IBD segments.

The Python script *variant_search.py* was used to generate horizontal bar plots, heatmaps and genetic frameworks for each of the variants that were investigated in this study.

3.3 IBD Tool Selection

There are many publicly available IBD detection tools that can be used to detect IBD segments from whole genome VCFs. From these, six of the most relevant and referenced tools in literature were selected. These include hap-IBD v1.0 (Zhou et al., 2020), RefinedIBD 17Jan20.102 version (Browning and Browning, 2013), FastSMC v.1.3.1 (Nait Saada et al., 2020), RaPID v.1.7 (Naseri

et al., 2019), Rapid-Query v.1.0 (Wei et al., 2023) and IBDSeq r1206 version (Browning and Browning, 2013). Evaluation of these tools was done with IBD Benchmark. Default parameters were used for the tools, except for RaPID and RaPID-Query which do not have any default parameters. For both of these, the parameters set by Tang et al. (2022) were used. The command line arguments to run the tools can be found in Appendix A.

3.4 IBD Benchmark

The selected tools were tested and validated using the IBD Benchmark tool created by Tang et al. (2022). This tool performs accuracy and power metrics to test the performance of IBD detection tools. IBD Benchmark requires four input files to calculate these performance measures. Three of these files need to be specified in the *Config.txt* file provided by the tool. These include the VCF file that was used as input to the IBD detection tool, the IBD output file generated from that tool, and a ground truth file which provides the ground truth IBD segments of the VCF. The fourth file is a genetic map of the chromosome. IBD Benchmark reads the configuration file and parses the input files on execution.

A readily available phased sequencing European dataset (closest to the Maltese population) of chromosome 20 of 4,000 individuals provided by Tang et al. (2022), and generated using msprime (Kelleher et al., 2016), was used as input. msprime simulates the ancestry and genetic diversity of individuals backwards in time by using certain population factors such as population size changes, migration and bottleneck effects. It takes into account LD, segregation and mutations, and generates a whole population tree for them. The ground truth IBD segments of this dataset were also provided. The IBD detection tools were tested on this dataset with 0%, 0.01% and 0.1% genotype error rates. The genotype errors were simulated using IBD Benchmark itself. Genotype errors in sequencing datasets with high quality scores generally do not exceed the value of 0.1%. The genotype quality (GQ) score is one such score which is provided in the Maltese VCF dataset used in this project. It is defined by $GQ = -10\log_{10}(\text{Error}$

Rate), where a GQ score of 10 is estimated to be around 10% genotype error rate, a score of 20 corresponds to an estimated 1%, and so on. Since most of our reads in the dataset have a GQ score of over 30, the estimated genotype error rate should be equivalent to or less than 0.1% (Wall et al., 2014).

IBD Benchmark uses different format parsers for the IBD output file, depending on the tool that generated the file. Five prebuilt format parsers for RaPID, FastSMC, hap-IBD, iLash and TPBWT are readily available with IBD Benchmark. For RaPID, FastSMC and hap-IBD, the respective parsers were used to parse the files. For RefinedIBD and IBDSeq, the format of the output file was modified using the Python scripts *refinedibd.py* and *ibdseq.py* to match hap-IBD's parser. However, since IBDSeq's output does not include a cM distance column for the detected IBD segments, the script also calculates them using the genetic map of GRCh37 chromosome 20. RaPID-Query's output file was modified using the Python script *query.py* to match RaPID's parser.

By default, IBD benchmark uses a 50% threshold, where a covered IBD segment is counted if a ground truth segment covers at least 50% of its length. This cutoff was changed through the tool's C# (Microsoft, 2024) files; the programming language used to develop the tool. Our thresholds varied from 50% to 100% at every 10% interval, as well as 95% and 99%. Mono compiler (Mono Project, 2024) was used to compile the C# files for each percentage threshold, as per the tool's documentation. The IBD detection tools were tested across all percentage thresholds and genotype error rates. When possible, the tools were run using their default parameters. A 2cM threshold was applied to all of the tools to prevent the inclusion of smaller false positive segments. Since RaPID and RaPID-Query do not have default parameter values, the parameters used by Tang et al. (2022) for RaPID were applied to both of these tools. Since IBDSeq requires unphased genomes, the partial phasing of the dataset was removed using the *sed* command, a stream editor command in Linux that performs text transformations on an input. It was used to replace all of the "|" symbols that represent phasing with "/" that represent unphasing. The parameters and command line arguments used to run the tools can be found in Appendix A.

Once IBD Benchmark results were generated for the six IBD detection tools, the results of the seven accuracy and power metrics were plotted in the form of graphs using the Python script *plots.py*. These provide a visualisation of the tools' performance at every percentage threshold and genotype error rate.

3.5 MAMI Study Dataset

In this study, 1,076 genomes extracted from the Maltese population that formed part of the Maltese Acute Myocardial Infarction (MAMI) project (Attard et al., 2014) were investigated for possible founder variants. These were made available by the MAMI study group as a single VCF of 840 GB in size. VCFs are the standard format used to store information about DNA polymorphisms, including SNPs, insertions and deletions, sometimes also with rich annotations. It allows for data from multiple genomes to be aggregated in a single VCF. As represented in Figure 3.2, a standard VCF consists of two sections, a header section and a data section. The header section starts by displaying several meta-information lines which start with the characters "##". These lines contain information about the descriptive tags and annotations used, date of creation, software used, reference sequence version and any other information related to the file's history. These are followed by a single TAB delimited line, comprising of a mandatory eight columns. This is marked by a single "#" character and contains data columns that represent the chromosome (CHROM), the baseposition of the variant (POS), the unique identifier of the variant (ID), the reference allele (REF), the alternate alleles (ALT), which can consist of multiple comma-separated characters, a Phred quality score (QUAL), site filtering information (FILTER), additional annotations (INFO) and FORMAT. The latter contains a colon separated field with information about the subsequent genotype columns. This can include information such as genotype alleles (GP), genotype quality (GQ) and read depth (DP), among others. This is followed by tab-delimited columns for every sample present in the data and will represent the genotype of the corresponding individual respective to each variant listed in the file. The genotypes are represented as 0 for the reference allele, 1 for the

first allele in the ALT column, 2 for the second allele in the ALT column, etc. A dot is used to represent any missing genotypes or values (Danecek et al., 2011).

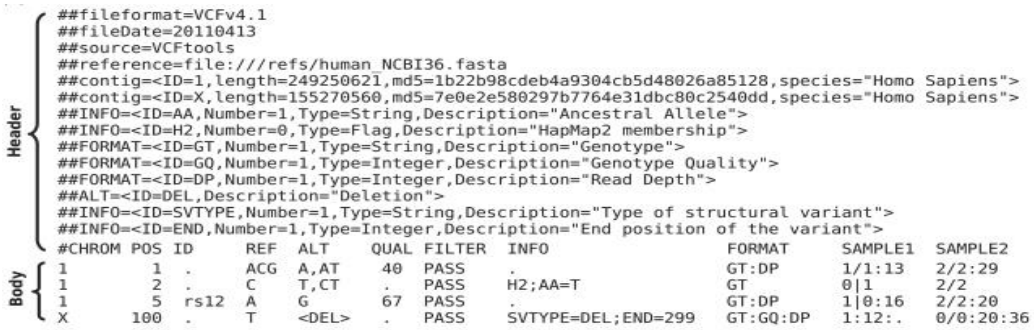

Figure 3.2: An example of a VCF showing meta-lines in the header section and samples in the body section of the file. Reproduced from Danecek et al. (2011) under the terms of the Creative Commons Attribution Non-Commercial License.

3.6 Haplotype Phasing

Also known as genotype phasing, haplotype phasing is a process of inferring haplotypes from genotype data and determining the parental origin of alleles as paternal and maternal, as in Figure 3.3. Genotypes obtained from sequencing are usually unphased, however phasing provides benefits in downstream analyses, including detection of deleterious compound heterozygotes, genotype imputation, genetic association testing, inference of population ancestry at a locus and detection of IBD segments. In the latter, haplotype phasing is particularly useful in the tracing of specific segments in multiple individuals that are inherited from a common ancestor, and in the identification of short segments characterized by rare variants (Hochreiter, 2013). Phased genotypes provide more accurate results and faster computational times, and this is why the majority of IBD detection tools require the use of phased data over unphased genomes (Browning and Browning, 2010).

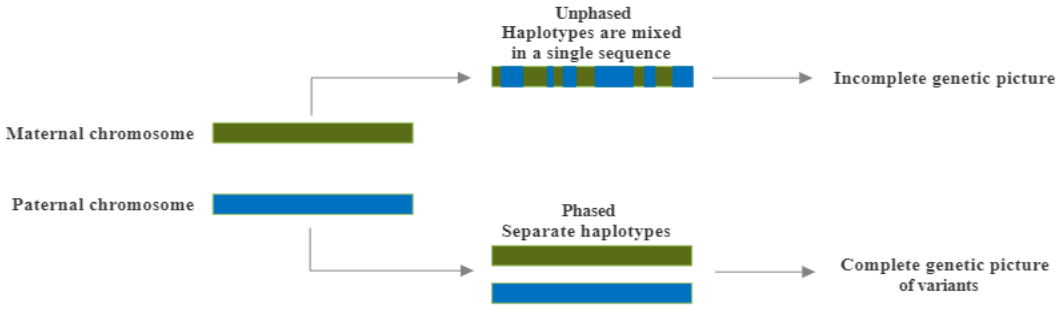

Figure 3.3: A simple visual representation of haplotype phasing. Phased haplotypes provide a better genetic picture of the variants.

The Maltese genomes were partially phased, meaning that phasing was only done on some of the variants and not the entire dataset. Since complete haplotype phasing of the entire dataset is a requirement of modern IBD detection tools, this was done beforehand with the use of the haplotype phasing tool Beagle v.5.4 (Browning et al., 2021). The haplotype phasing method present in Beagle is designed to handle large sequence datasets which can contain hundreds of millions of markers. It applies a two-stage algorithm which uses Li and Stephens (2003) HMM to limit the amount of data that is stored in memory, hence improving computational efficiency. The algorithm also allows for multithreading, making the programme more memory efficient. The first part of phasing involves high-frequency markers, where a progressive phasing method is used to incrementally expand the set of phased heterozygotes. Each heterozygote's genotype is set to either "finished" or "in progress". At the start of the phasing progress, all of the heterozygotes are set to in progress. The programme then iterates through each heterozygote and the phase is estimated and updated with reference to the previous heterozygote. Those heterozygotes that are most confidently phased are set to finished, and can no longer be changed by further iterations (Browning et al., 2021). This ratio of confidence is calculated using the forward-backward HMM algorithm (Rabiner, 1989) by multiplying the probabilities of the two haplotypes in a diplotype. Assuming Hardy-Weinberg equilibrium, the updated phase of the target heterozygote in relation to the prior heterozygote is determined by the diplotype with the higher probability. Confidence in the inferred phase is expressed as the ratio of the bigger diplotype probability to the smaller diplotype probability. This is followed by the second stage of phasing high-frequency markers as a haplotype scaffold for allele imputation, and this allows the phasing of low-frequency markers from inferred allele probabilities (Browning et al., 2021).

Figure 3.4 shows the difference between a phased and an unphased VCF. In an unphased dataset the haplotypes are separated by a "/" character whereas in a phased dataset they are separated by a "|" character. The partial phasing of the dataset was removed using the *sed* command to replace all of the "|" symbols with "/". This was followed by the filtering of 232 samples from the 1,076 in the

dataset, relatives of the other 844 individuals. Since IBD segments are inherited from the parents to the offspring, removal of relatives would ensure detection of true IBD segments in unrelated individuals. A file with relationship information of the MAMI study participants was provided by the MAMI study group (Attard et al., 2014). The remaining 844 samples were used for IBD detection.

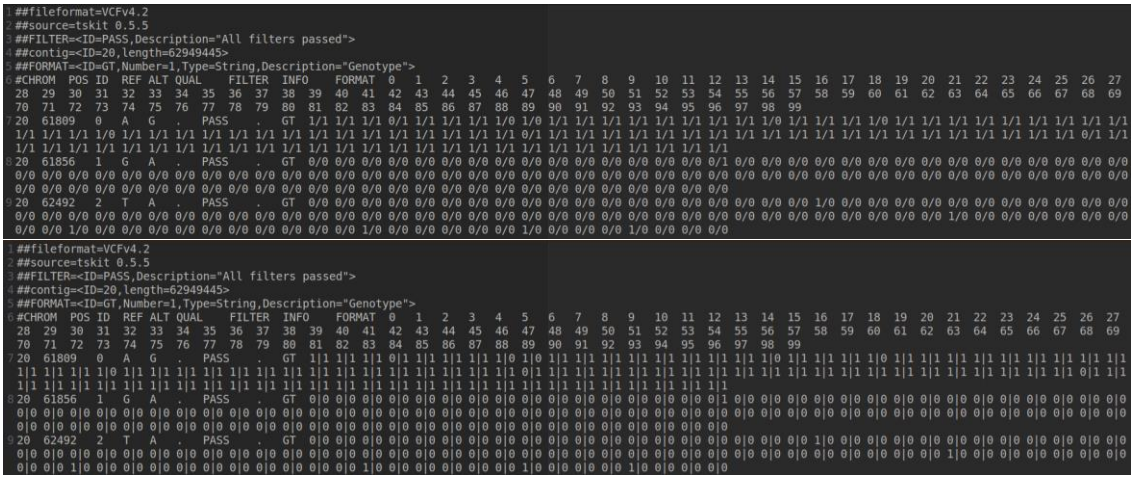

Figure 3.4: A representation of the same VCF in the unphased (top) and phased (bottom) versions.

The unphased and filtered VCF was then divided by chromosome with the use of VCFtools v.0.1.17. Once complete, phasing with Beagle was done. The input VCF is specified by the *'gt'* parameter, whereas the output VCF is specified by the *'out'* parameter. The option *'map'* parameter is used to specify Plink genetic maps, according to the chromosome being phased. Beagle uses such maps to estimate the genetic distance between positions. If no map is specified, Beagle assumes a recombination rate of 1cM per Mb. The *'nthreads'* parameter specifies the number of allocated computational threads, whereas the *'window'* parameter specifies the cM distance of each sliding window (default set to 40.0).

The command line arguments used to perform dataset relative filtering, chromosomal extraction and haplotype phasing can be found in Appendix A.

3.7 Variants Under Study

The compilation of a list of variants consisted of variants that were identified in a number of Maltese patients with rare conditions (Table 3.1). They

are related to a range of disorders, which include neurological, reproductive and endocrine, haemoglobin, kidney, metabolic and innate disorders.

Table 3.1: The list of variants that were investigated in this study, including their variant identifier, genetic locus, affected gene and nucleotide and predicted protein changes. N/A indicates that the variant was never found in that population. A value of 0.000 indicates that the MAF of the variant is very close to 0, but on rare occasions may have been found in the population. Abbreviations: MAF: minor allele frequency, MAL: Maltese, NFE: Non-Finnish Europeans, AFR: African, PD Parkinson's disease, IHH Idiopathic hypogonadotropic hypogonadism, HPFH hereditary persistence of foetal haemoglobin, SR Sepiapterin reductase, DHPR Dihydropteridine reductase, CAAHD Congenital arthrogryposis with anterior horn cell disease, LCCS Lethal congenital contracture syndrome 1, NS Nephrotic syndrome.

Variant Identifier	Gene	Associated condition/phenotype	Chr.	Basepair position	DNA Nucleotide Change	Predicted Protein Change	MAF_MAL	MAF_NFE	MAF_AFR
rs66812916	*CDCP2*	PARK10 locus; PD	1	54,139,647	G>GT	p.P408RfsX46	0.008	0.000	0.000
rs71745629	*KISS1*	Fertility/IHH	1	204,190,483	CT>C	p.X139fs	0.255	0.214	0.095
rs4889	*KISS1*	Fertility/IHH	1	204,190,659	G>C	p.P81R	0.298	0.252	0.372
rs35431622	*KISS1*	Fertility/IHH	1	204,190,794	T>C	p.Q36R	0.063	0.051	0.204
rs398122922	*SPR*	SR deficiency	2	72,891,345	A>G	r.spl	0.015	0.000	0.000
rs104893836	*GNRHR*	IHH	4	67,754,019	T>C	p.Q106R	0.017	0.004	0.001
rs104893863	*QDPR*	DHPR deficiency	4	17,511,987	C>T	p.G23D	0.007	0.000	0.000
rs2276973	*TACR3*	IHH	4	103,656,225	T>C	p.K286R	0.008	0.001	0.024
rs1564162129	*GLE1*	CAAHD/LCCS	9	128,541,151	C>T	p.S693F	0.003	N/A	N/A
rs35553496	*HBB*	Haemoglobin Valletta	11	5,226,630	T>G	p.T88P	0.015	0.000	0.000
rs36049074	*HBG2*	Haemoglobin F-Malta 1	11	5,253,368	T>C	p.H118R	0.014	N/A	N/A
rs753540084	*LRRK2*	PD	12	40,274,905	A>G	p.N618S	0.019	N/A	N/A
rs770541847	*KISS1R*	IHH	19	919,929-919,965	GCGCGCCTACTGCAGTGAGGCCTTCCCCAGC>G	p.Y190_A199del	0.001	0.000	0.000
rs267607202	*KLF1*	HPFH	19	12,885,368	T>A	p.K288X	0.001	N/A	N/A
rs267606919	*NPHS1*	NS	19	35,831,056	G>A	p.R1160X	0.019	0.000	0.000

3.8 Investigating the Founder Status of Variants

RaPID, the best performing IBD detection tool, was used for the remainder of the project to generate IBD segments and from there investigate the variants and identify which are founder variants of the Maltese population. After sample filtering, chromosome extraction and haplotype phasing of the dataset was performed, IBD detection was performed with RaPID's optimal parameters. These include a window size (-w) of 5, 10 number of runs (-r) and 2 number of successes (-s). The phased chromosomal files were gzipped, as per the RaPID's requirements. Along with directory sorting, these steps were done with the use of the bash script *generate_ibds.sh*. IBD detection of chromosome Y was not possible as genetic maps for this chromosome do not exist

IBD segment detection was performed on a chromosomal basis with a minimum cM threshold of 2, as specified in the tool's '-d' parameter. Most IBD tools choose this as the default value since segments inherited from common ancestors from 500 to 1,500 years ago are longer than 4 cM, whereas segments from the last 500 years are longer than 10 cM. Only a small number of segments longer than 2 cM are inherited from ancestors longer than 4,000 years ago (Ralph and Coop, 2013). IBD detection with RaPID was also performed with a minimum threshold of 0.5 cM, and the results of both thresholds was compared.

3.8.1 Generating Figures and Results

The script *variant_search.py* is used to generate outputs to determine the founder status of variants. The first step of the script is to read the "*variants.txt*" file, which consists of a list of variants to be investigated. This includes the chromosome number, basepair position, reference allele and alternate allele of the variants in tab-delimited format. Using this information, the script goes to the respective chromosomal VCF and variant position to extract the sample identifier of the individuals that have the variant, as well as the genotype data of such individuals. It also goes to the respective IBD file containing all of the identified segments of that chromosome, and extracts the segments that pass through the

variant basepair position. This information includes the start and end basepair positions of the segment, as well as a pair of sample identifiers in which this segment was identified. At least one of the individuals in this pair needs to have the variant for the segment to be extracted. This condition ensures that the extracted IBD segments are of the variant, and eliminates any other overlapping segments that the variant does not form part of.

Using the start and end basepair positions of the segments, a horizontal bar plot is plotted for every segment and the respective pair of individuals. The common range of the identified IBD segment, which is found among all of the individuals in the generated horizontal bar plot, is highlighted in orange. The plot also highlights any ROH segments, which will be reported. The common range of positions is then plotted in the form of a heatmap, showing the genotype sequence of all 844 unrelated individuals in the dataset. The individuals are grouped together into homozygous alternate, heterozygous and homozygous reference individuals for the variant being investigated. This will highlight the difference in genotype sequences between the three groups and highlights IBD segments of the variant which should only be found in individuals with the variant (therefore not in homozygous reference individuals). The script also has the option of generating a smaller heatmap, focusing on the basepair positions around the variant and limited to a random 50 homozygous reference individuals (chosen by equal probability sampling) and all individuals with the variant (heterozygous and homozygous alternate individuals). This allows the user to visualise better the segments that go through the variant's position, and the option can be turned on and off by the user. Any positions that are of the homozygous reference sequence among all individuals are filtered out of the heatmaps as we are mainly interested in the variants and their alternate alleles.

At some basepair positions throughout the VCFs, more than one variant is sometimes presented. This is caused by indel calling in low complexity regions. These regions have highly repetitive sequences, which lead to misalignment of the reads to the reference sequence. While some of the reads are correctly aligned, the misaligned reads will produce fragments of the original indel, which

are incorrect and do not really exist in the genome. This generates many false positives and false negatives, and also impairs annotations in downstream analysis (Fang et al., 2014; Gong et al., 2024). Therefore, it is important to identify the true reads. At these positions, filtering steps were taken to identify the correct indel variant. This was done by reading through the respective Binary Alignment Map files. This type of file is a binary compressed version of the Sequence Alignment Map file which contains sequences (up to 128Mbp) generated from next generation sequencing technologies and aligned to a reference sequence (Li et al., 2009). The script uses these files to read the aligned sequences and select the indel with the most reads, which would be the correct indel.

The IBD framework is then plotted using the data in this range of positions, by taking the allele of each individual that contains the variant and finding the most common allele sequence between the individuals. This ensures that only the alleles on the same chromosome as the variant are taken into consideration, meaning that the framework is of the variant allele. For an allele to be included in the framework, it must be present in at least 90% of the individuals with the variant being investigated. This makes the framework similar among the individuals that have the same IBD segment and variant, therefore suggesting that the variant being investigated is indeed a founder variant. The reference sequence of these positions is also plotted for comparison. Any alleles on the variant's chromosome that are equal to the reference sequence are filtered out as these do not give any additional information to the framework. The framework and respective reference sequences are written to a text file '*variant_framework.txt*'.

Altogether, these different outputs and their interpretations will be used to identify whether IBD segment detection can be used to classify whether a variant is a founder variant or not. All of the aforementioned implementations were used to design a bioinformatics pipeline which facilitates the identification of founder variants within IBD segments.

3.9 Summary

In this chapter, the use of IBD Benchmark to choose the overall best tool was described. A description of the Maltese dataset was given, with the involved preprocessing steps that include dataset filtering, chromosomal extraction and haplotype phasing. This was followed by the list of variants to be investigated in the Maltese population for founder status. Finally, a description of the scripts used to generate horizontal bar plots, heatmaps and variant frameworks for founder variant analysis were described.

Ethical approval to carry out this study was obtained from the Centre of Molecular Medicine and Biobanking Research Ethics Committee, with the ethics application I.D. of CMMB-2023-00005.

4. Results and Discussion

The six IBD detection tools hap-IBD, RaPID, RaPID-Query, RefinedIBD, IBDSeq and FastSMC were tested for accuracy and power, with the best performing tool chosen to generate IBD segments and investigate whether they can be used to determine the founder status of variants in the Maltese. IBD Benchmark (Tang et al., 2022) was used to test the tools for power and accuracy across different thresholds and genotype error rates, using Tang et al.'s readily phased sequencing European dataset of 4,000 individuals. The best performing tool was optimised and used for the remainder of the study to generate IBD segments of the Maltese dataset. Generated horizontal bar plots, heatmaps and variant framework sequences will also be presented, with the aim of identifying the founder status of variants in the Maltese. Identification of such founder variants will allow for a more specific approach in genetic testing of the Maltese population and the development of preventive measures and treatments.

4.1 Accuracy and Power of IBD Detection Tools

RaPID, RaPID-Query, hap-IBD, RefinedIBD, IBDSeq and FastSMC were the most commonly used IBD detection tools in literature, and their performance was tested using IBD Benchmark's seven measures. These include accuracy, length accuracy, length discrepancy, recall, power, accumulative recall and accumulative power. The most notable metrics are accuracy and recall, as these directly measure the tool's ability to detect ground truth IBDs. Length discrepancy and length accuracy are also important as these measure the tool's ability to detect the entirety of the segment. IBD Benchmark calculates the metrics across different cM bins to test the tool's ability across different cM lengths as follows: (2, 3), (3, 4), (4, 5), (5, 6), (6, 7) and (7, ∞) cM bins. By default, IBD Benchmark uses a 50% threshold to calculate these measures, therefore a covered IBD segment is counted if a ground truth segment covers at least 50% of its length (Tang et al., 2022). The tools were tested across multiple percentage benchmark thresholds ranging from 50% to 100% and different genotype error

rates of 0%, 0.01% and 0.1%. The latter are generated through the genotyping process, starting from sampling and DNA extraction to molecular and data analysis, as well as chance and human error. This can cause a reduction in detection power in downstream analysis, and should be accounted for. In the subsequent section, IBD benchmark results will be presented and compared to those produced by Tang et al. (2022). In their publication, Tang et al. (2022) tested five IBD detection tools, which include RaPID, hap-IBD, FastSMC, TBPWT and iLash. We repeated testing on RaPID, hap-IBD and FastSMC, and added RaPID-Query, RefinedIBD and IBDSeq which have not been tested yet, since these are the six most referenced tools in literature. Therefore, the comparison will be limited to the tools that we tested and to the 50% threshold since higher benchmarking thresholds were not tested by Tang et al. (2022).

4.1.1 Genotype Error Rate: 0%

Figure 4.1 shows the different metrics that were used to test the tools at 50% threshold and no genotype error rate. Except for IBDSeq, the tools show high accuracy and length accuracy across all cM bins. IBDSeq and RefinedIBD expressed high length discrepancy when compared to the other tools, especially with segments larger than 7cM. The best performing tools, RaPID and FastSMC, had high recall and power values, with IBDSeq showing the lowest performance overall. These results are comparable to Tang et al.'s (2022) results while testing out RaPID, hap-IBD and FastSMC at the default 50% threshold and no genotype error rates. RaPID is the best performing tool overall.

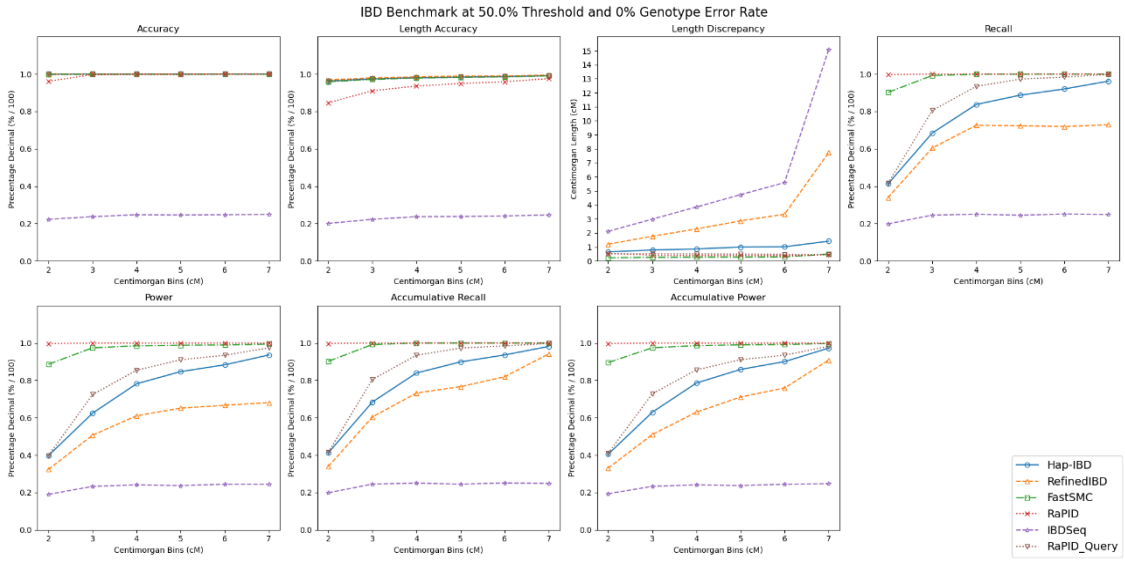

Figure 4.1: The seven different accuracy and power metrics in IBD Benchmark developed by Tang et al. (2022) to test IBD detection tools, at 50% threshold and no genotype error rate. RaPID and FastSMC are the overall best performing tools.

As the percentage benchmark threshold increases, there is an overall slight decrease in performance across all of the tools up to 90% threshold as it becomes increasingly difficult to identify larger parts of IBD segments. RaPID was the most affected, with a decrease in accuracy at the lower cM bins, with a minor decrease for the other tools (Figure 4.2). FastSMC and RaPID retained good recall measures, with RefinedIBD experiencing the largest decrease in performance. Overall, RaPID and FastSMC are the best performing tools up to the 90% threshold as they performed the best across most of the metrics.

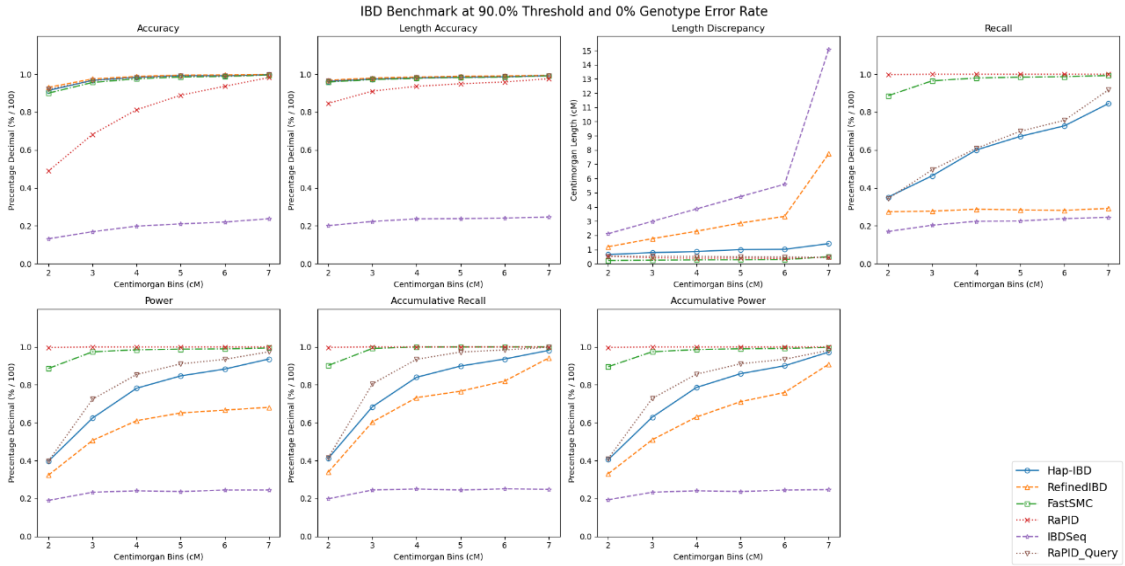

Figure 4.2: IBD Benchmark results of the IBD detection tools at 90% threshold and 0% genotype error rate. RaPID's accuracy decreased with a higher percentage threshold, but maintained good results across the other metrics. RaPID and FastSMC are the best performing tools overall.

At 99% threshold (Figure 4.3), all of the tools experience a large decrease in accuracy, with RefinedIBD showing the highest value while RaPID and IBDSeq showing the lowest. However, RaPID was the most robust in terms of recall, still showing near perfect values at this threshold. The benchmark threshold did not seem to have a large effect on the length accuracy and power of the tools. At the 99% threshold, RefinedIBD is the best tool in terms of accuracy while RaPID is the best tool in terms of recall, power and length discrepancy.

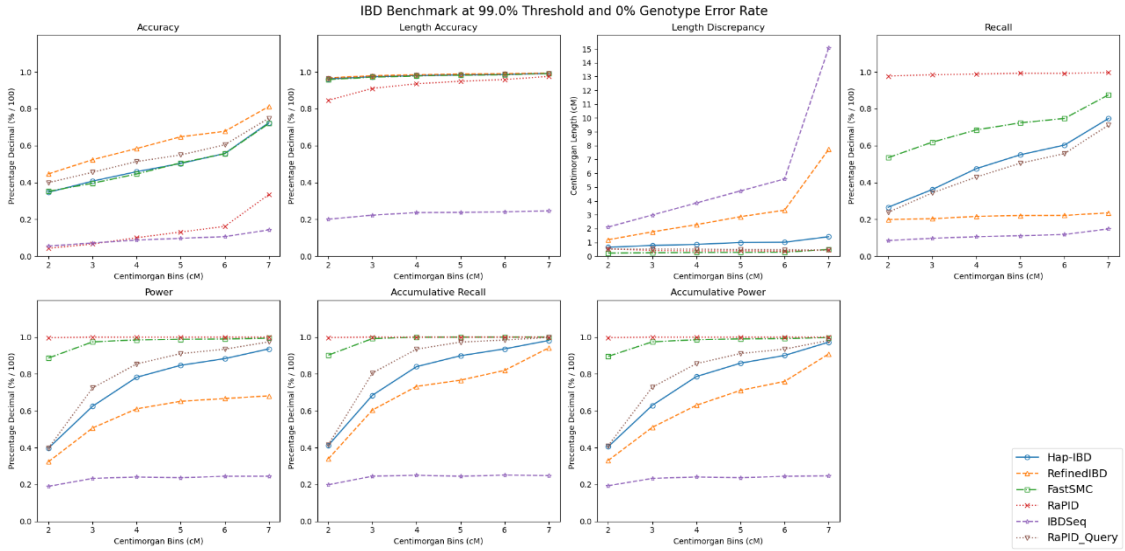

Figure 4.3: IBD Benchmark results of the IBD detection tools at 99% threshold and 0% genotype error rate. RefinedIBD has the best accuracy. RaPID has the best recall, power and length discrepancy.

When the tools were tested to identify the entire IBD segments (100% threshold, Figure 4.4), they all showed low accuracy and recall values, except for RaPID which maintained high recall. This makes RaPID the best performing tool.

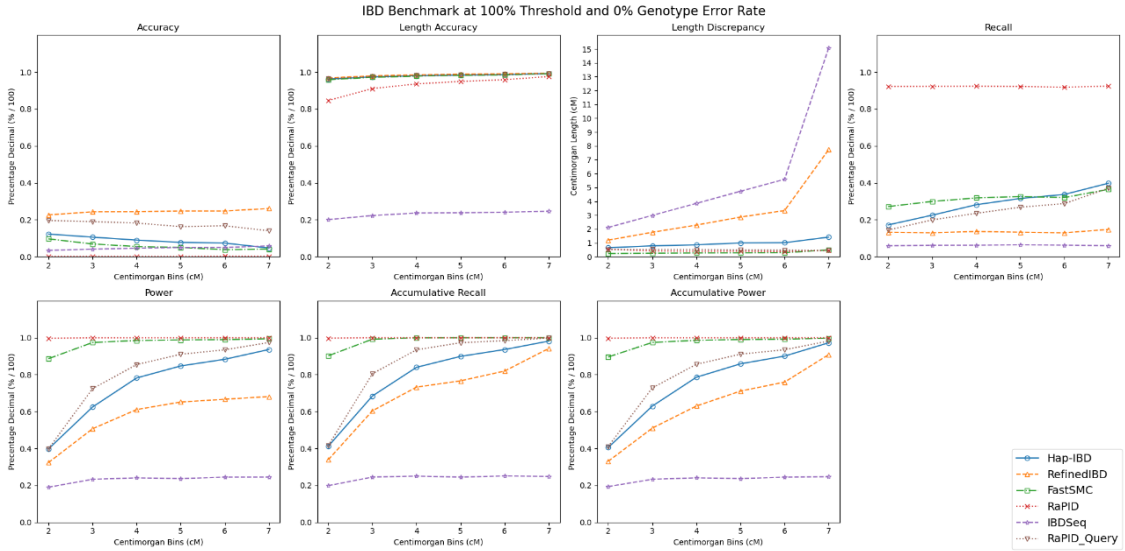

Figure 4.4: IBD Benchmark results of the IBD detection tools at 100% threshold and 0% genotype error rate.

Overall, the results show that the performance of the tools generally decreases as the threshold increases, while the genotype error rate is kept constant. This is because it becomes increasingly difficult for the tools to identify larger parts of IBD segments or the entirety of them due to the different genetic diversity between individuals in a population. Based on the results it also seems easier to detect the larger IBD segments as these would be more obvious.

With 0% genotype error rate, we can conclude that hap-IBD, FastSMC, RefinedIBD and Rapid-Query are able to maintain high accuracy up to the 90% threshold. FastSMC also maintained a high recall value up to this threshold, however RaPID remained the most robust across all thresholds. RaPID and FastSMC expressed the highest power values overall, followed by Rapid-Query, hap-IBD, RefinedIBD and IBDSeq respectively. Except for IBDSeq, all of the tools maintained high length accuracy across all of the thresholds.

However, real data always contains some degree of genotype error rate. High throughput sequencing datasets with high quality scores, such as the one being used in this study, generally do not exceed a genotype error rate of 0.1%. Therefore, in the subsequent subsection the effect of genotype error rate on the tools will be investigated at 0.01% and 0.1% genotype error rates.

4.1.2 Genotype Error Rates: 0.01% and 0.1%

Figure 4.5 shows the effect of the introduction of 0.01% genotype error rate on the performance of the tools at the 50% threshold. This does not seem to have an effect on the accuracy and length accuracy of the tools when compared to 0% genotype error rate. FastSMC and hap-IBD show higher length discrepancy, especially at the higher cM bins. Except for RaPID, the tools experienced a negative effect in performance in power and recall. At the 50% threshold and 0.01% genotype error rate RaPID is the best performing tool.

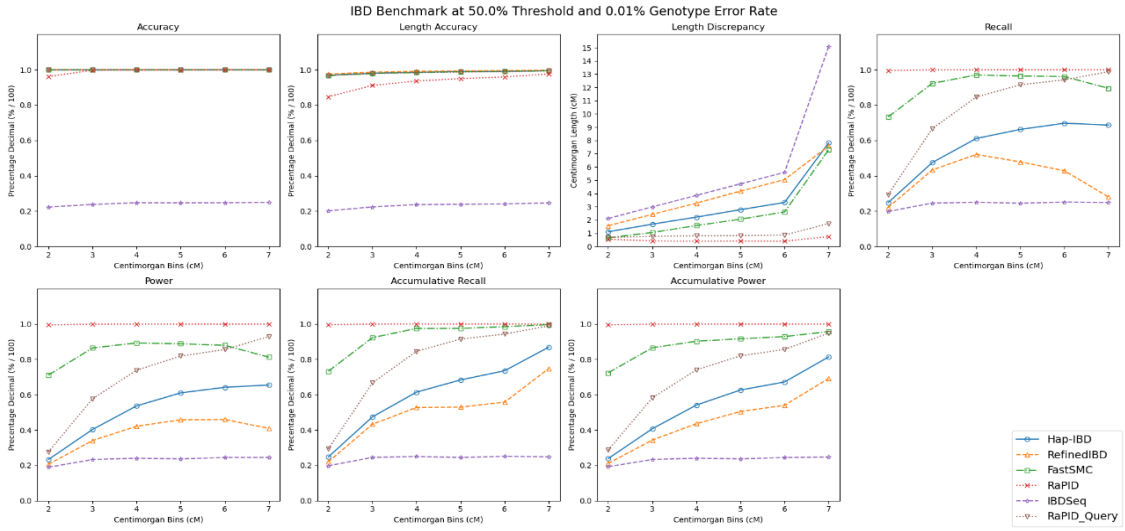

Figure 4.5: IBD Benchmark results of the six IBD detection tools with the introduction of 0.01% genotype error rate at 50% threshold. This has a slight negative effect on the performance of the tools. RaPID is the best performing tool overall.

The negative effect of genotype error can be seen to increase with the introduction of 0.1% genotype error rate (Figure 4.6). While the accuracy and length accuracy of the tools was unaffected, the tools experienced a negative effect in length discrepancy, recall, power, accumulative recall and accumulative power. RaPID was the most robust tool, maintaining the highest performance. This shows that genotype errors have a significant impact on detecting IBD segments as true genotypes can be altered and may disrupt such segments. This can cause true IBD segments to be missed (false negatives), unless the IBD detection tool is robust enough to cater for the genotype error. These findings are comparable to Tang et al's (2022) sequencing data results.

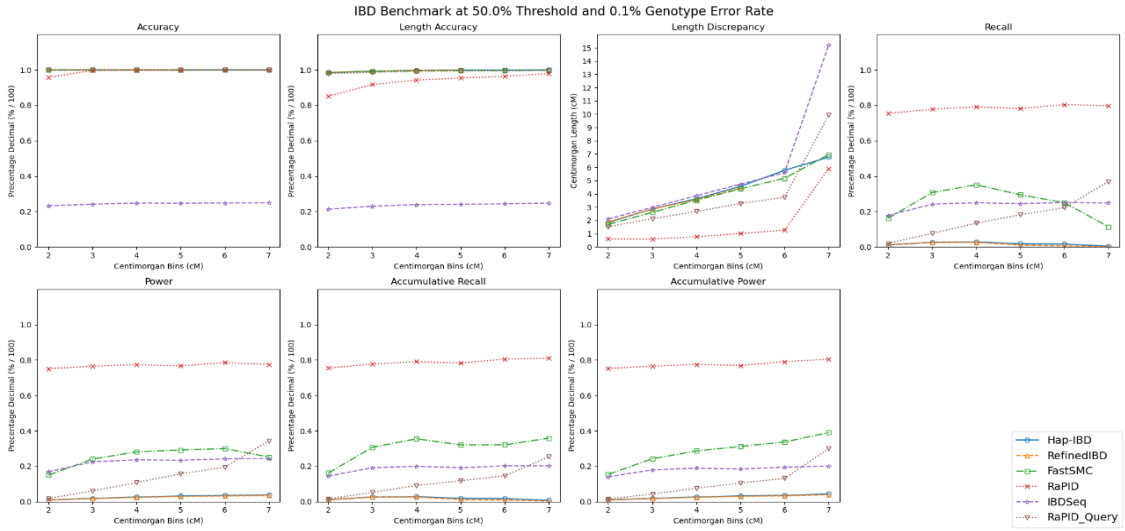

Figure 4.6: IBD Benchmark of the IBD detection tools with the introduction of 0.1% genotype error rate at 50% threshold. RaPID stands as the best performing tool.

The tools experience a gradual slight decrease in performance with every threshold interval, up till the 90% threshold, as it becomes increasingly difficult to identify larger parts of IBD segments. Figures 4.7 and 4.8 show the performance of the tools at 90% threshold and 0.01% and 0.1% genotype error rates respectively. RaPID and IBDSeq show a decrease in performance in terms of accuracy and length accuracy at this threshold. The rest of the tools show a decrease in length discrepancy, recall and power metrics.

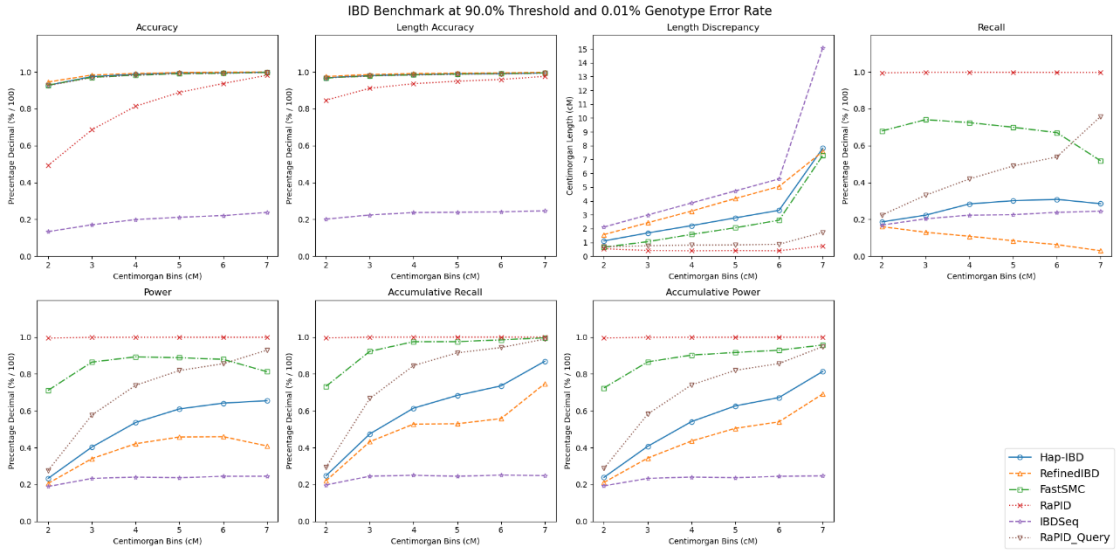

Figure 4.7: IBD Benchmark results of the IBD detection tools at 90% threshold and 0.01% genotype error rate. RaPID has the best performance in length discrepancy, recall, power, accumulative recall and accumulative power. RefinedIBD is the best in terms of accuracy and length accuracy.

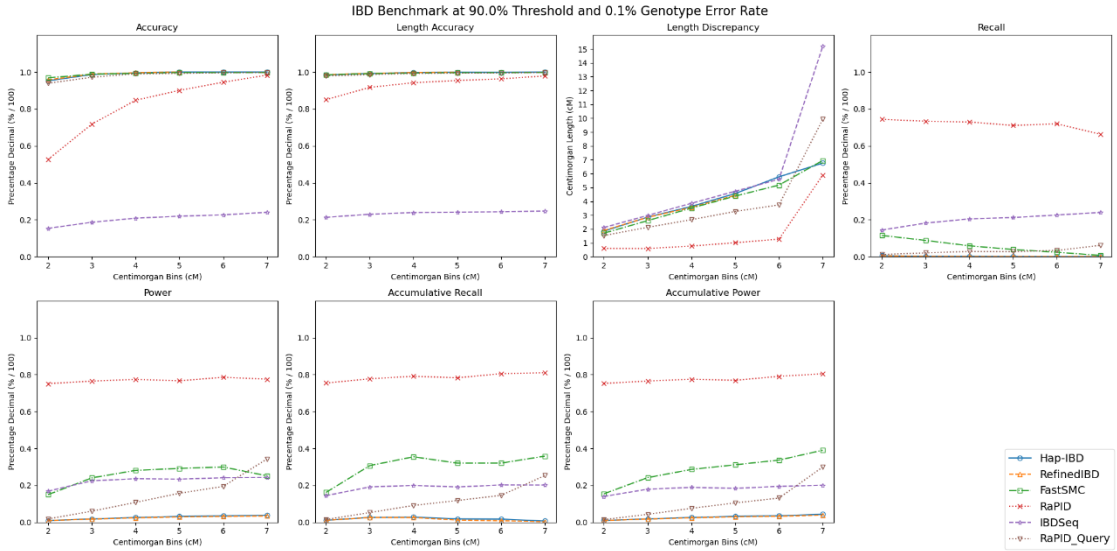

Figure 4.8: IBD Benchmark results of the IBD detection tools at 90% threshold and 0.1% genotype error rate. RaPID has the best performance in length discrepancy, recall, power, accumulative recall and accumulative power. RefinedIBD is the best in terms of accuracy and length accuracy.

The 99% threshold has a major effect on the performance of the tools, as shown in Figures 4.9 and 4.10. With 0.01% genotype error rate, the tools experience major decrease in accuracy and recall. RaPID and IBDSeq have the lowest accuracy values. RaPID maintained high recall while the rest of the tools showed a major decrease. Similarly, there was a decrease in performance with 0.1% genotype error. Compared to the 90% threshold, there was a decrease in accuracy, recall and power across all of the tools. RaPID and IBDSeq were the lowest in accuracy, however RaPID maintained high recall and power and RefinedIBD maintained high accuracy when compared to the rest of the tools.

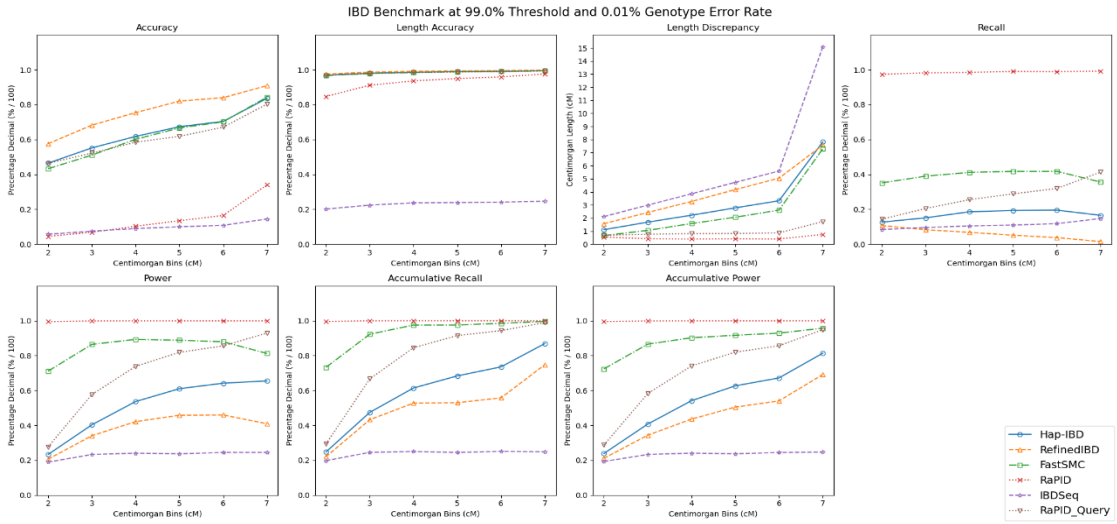

Figure 4.9: IBD Benchmark results of the IBD detection tools at 99% threshold and 0.01% genotype error rate. RaPID has the best performance in length discrepancy, recall, power, accumulative recall and accumulative power. RefinedIBD is the best in terms of accuracy and length accuracy.

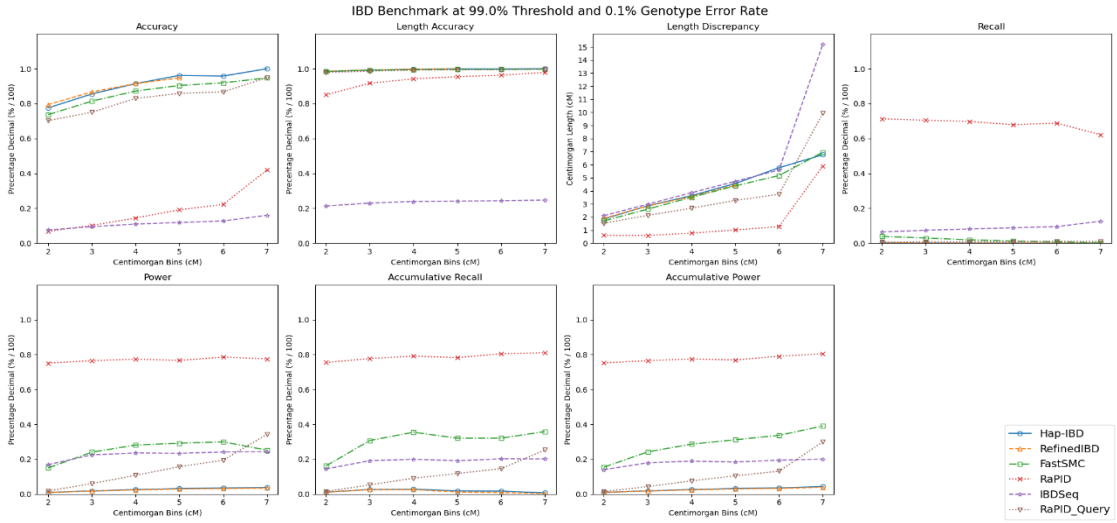

Figure 4.10: IBD Benchmark results of the IBD detection tools at 99% threshold and 0.1% genotype error rate. RaPID has the best performance in length discrepancy, recall, power, accumulative recall and accumulative power. RefinedIBD is the best in terms of accuracy and length accuracy.

When the tools were tested to identify 100% of the IBD segments at 0.01% genotype error rate (Figure 4.11), RefinedIBD showed the highest accuracy, whereas RaPID showed the highest recall and power values. With 0.1% genotype error rate (Figure 4.12), the tools showed a decrease in performance when compared to 0.01% error rate, except for the accuracy metric where hap-IBD, FastSMC, RefinedIBD and RaPID-Query showed higher accuracy values. RaPID still remained the best tool in terms of recall and power, significantly outperforming the rest of the tools. Naturally, all of the tools experienced a decrease in performance when compared to the lower 99% threshold.

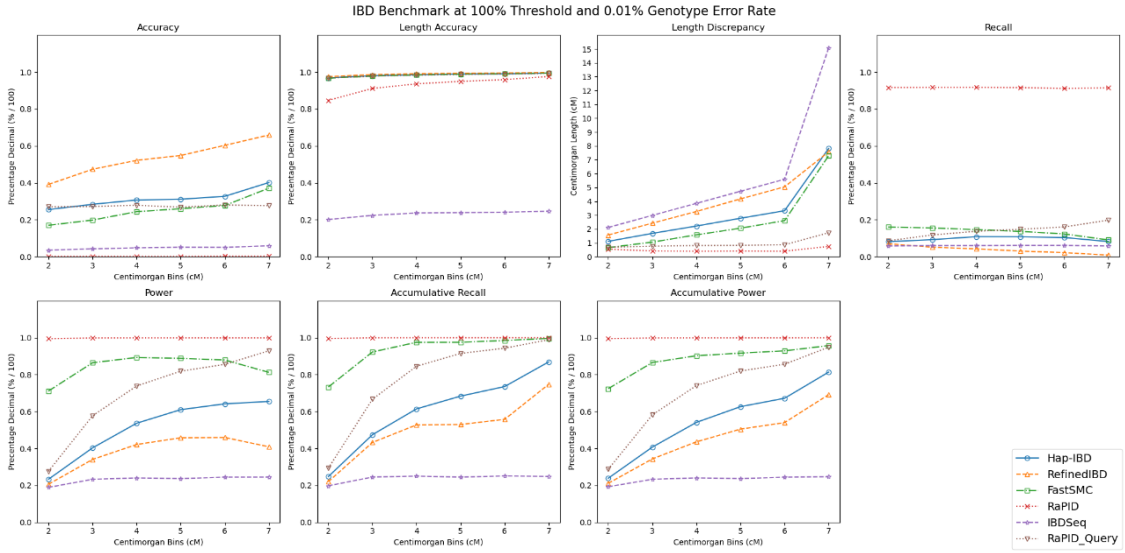

Figure 4.11: IBD Benchmark results of the IBD detection tools at 100% threshold and 0.01% genotype error rate. RaPID has the best performance in length discrepancy, recall, power, accumulative recall and accumulative power. RefinedIBD is the best in terms of accuracy and length accuracy.

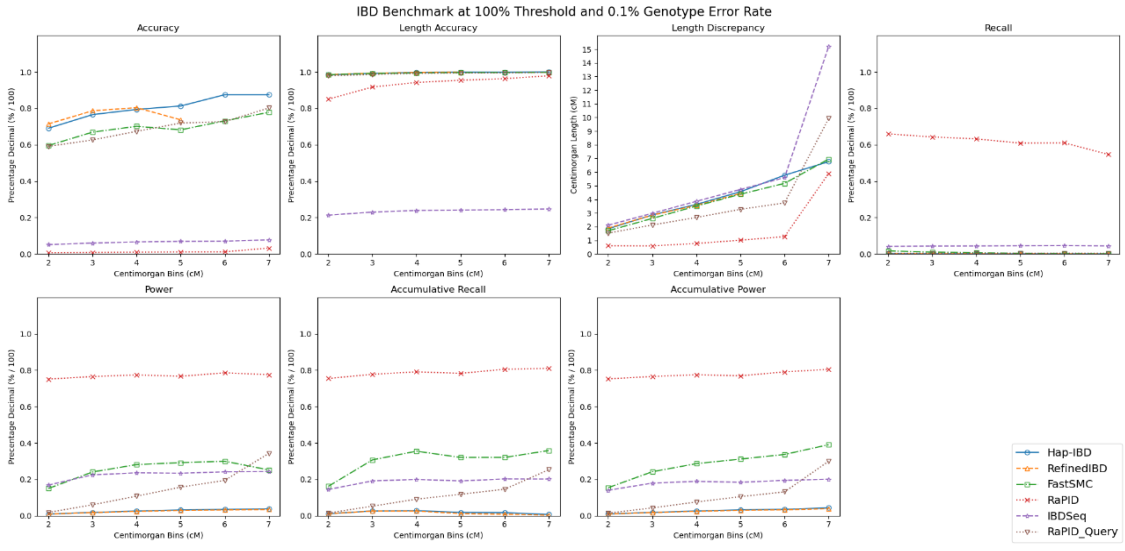

Figure 4.12: IBD Benchmark results of the IBD detection tools at 100% threshold and 0.1% genotype error rate. RaPID has the best performance in length discrepancy, recall, power, accumulative recall and accumulative power. Hap-IBD is the best in terms of accuracy and length accuracy.

Thus, the performance of the IBD detection tools decreases as the benchmarking threshold increases. This is because it becomes increasingly difficult to identify larger parts of IBD segments as these may be affected by recombination events that can break the IBD segment into smaller segments. The performance also decreases with an increase in the genotype error rate. Genotype errors can alter the genotype sequence and generate false negative segments instead of the true IBD segments. Unless the IBD detection tool is able to cater for such errors, it will not be able to identify these segments and thus its performance decreases. As with the thresholds, it is easier to detect the longer IBD segments. Up until the 90% threshold, Rapid-Query, RefinedIBD, hap-IBD and FastSMC maintained high accuracy at the 0.01% and 0.1% genotype error rates. RaPID had a decrease in accuracy when detecting the smaller segments (less than 6cM in size). This was not the case in recall and power as only RaPID managed to maintain high performance in these metrics, even at 100% threshold. RaPID also had the least length discrepancy overall. IBDSeq's performance was low across all of the metrics. With RaPID outperforming the other tools in most of the metrics offered by IBD Benchmark, it was chosen as the best performing tool, followed by FastSMC. The parameters of both of these tools were tested further, with the aim of increasing their performance across the metrics.

4.1.3 Tool Wall-Clock Run Time

Table 4.1 represents the wall-clock run time of the six IBD detection tools when detecting IBD segments on the test European dataset of 4,000 individuals using default parameters. The wall clock time represents the amount of time that a process takes in real world time. These were run on a server having 1 TB of RAM and the time was measured using the *time* command in linux. hap-IBD was the most efficient tool, followed by RaPID, RaPID-Query, RefinedIBD, FastSMC and IBDSeq in that order. The processing time of RaPID, RaPID-Query, hap-IBD and FastSMC generally decreases as the genotype error rate increases, whereas RefinedIBD's and IBDSeq's processing time increases. With RaPID and FastSMC being the best performing tools, RaPID is much more efficient than FastSMC and

is therefore the preferred tool. Tang et al.'s (2022) findings also show hap-IBD as the most efficient tool, followed by iLash, RaPID and TPBWT. However direct comparison of the results cannot be made since different datasets were used to test the tools' run time.

Table 4.1: The wall clock time of IBD detection tools across different genotype error rates.

Tool / Genotype Error Rate	0% Genotype Error	0.01% Genotype Error	0.1% Genotype Error
RaPID	3min. 50s.	3min. 49s	3min. 9s
FastSMC	8hrs. 35min.	9hrs. 24min.	5hrs. 39min.
hap-IBD	1min. 54s	1 min. 55s.	1min. 32s
RefinedIBD	1hr. 56min.	1hr. 57min.	2hrs. 4min.
IBDSeq	9hrs.	15hrs. 21min.	15hrs. 52min.
RaPID-Query	1hr. 45min.	1hr. 45min.	1hr. 36min.

4.2 Determining RaPID's Parameters

With RaPID being the best performing tool, its parameters were tested further in IBD Benchmark by giving them different values. These include the number of runs (*r*), the number of successes (*s*, minimum number of runs required for a match to be taken into consideration) and the window size (*w*, number of SNPs per window). The tool does not have any default values for these parameters. In the first part of this task, the window sizes of the tool were tested, with the values of 1, 3, 5 and 30. The values correspond to the margin of error that is allowed by the tool, with higher window size values allowing a higher margin of error (Seidman et al., 2020). The other two parameters, number of runs and number of successes were kept constant at the values of 10 and 2 respectively. These were tested across the different thresholds from 50% up to 100%, as well as the genotype error rates of 0%, 0.01% and 0.1%. Since FastSMC was the second-best performing tool, its results were also compared. FastSMC's optional parameters do not affect the quality of the results but primarily handle

data organisation and output format, and thus were kept as default. Plots for the different window sizes were generated using the Python script *rapid_windows.py*.

4.2.1 Window Sizes at 0% Genotype Error Rate

RaPID's window sizes were first tested at the 0% genotype error rate. Figure 4.13 represents the 50% threshold. The accuracy and length accuracy were high across the different window sizes. The recall and power metrics were lowest at the window size of 30, with the rest of the windows showing high values. This was also the case in the length discrepancy parameter, which was higher at the window size of 30.

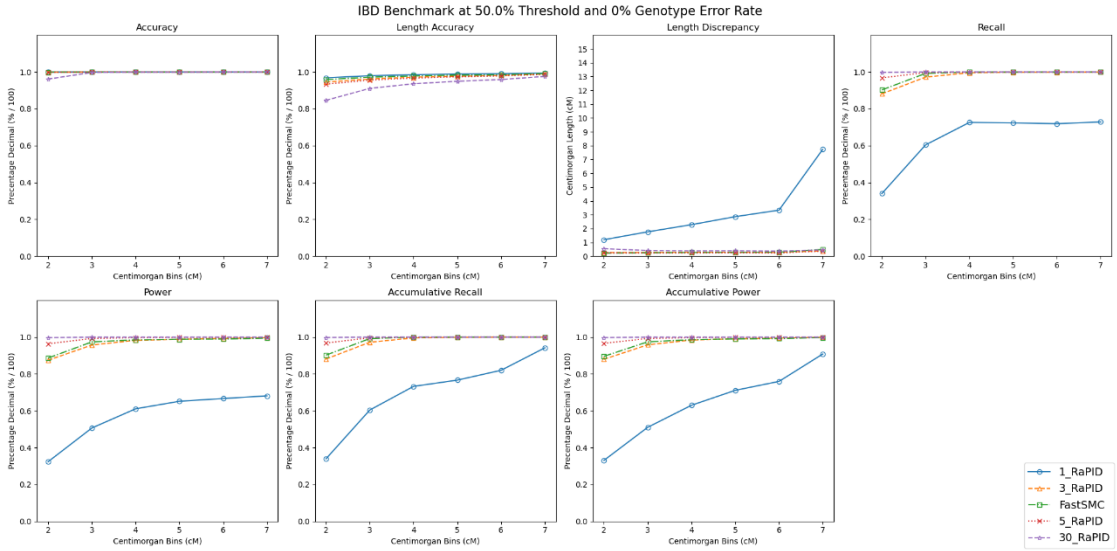

Figure 4.13: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 50% threshold and 0% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power.

As the threshold percentage increases, the performance of the tool decreases, as in Figure 4.14 at 95% threshold. There is an overall decrease in performance across all metrics, with the length accuracy, length discrepancy, recall and power values experiencing the least change. The window size of 1 shows the best accuracy and length accuracy while that of 30 shows the best recall and length discrepancy.

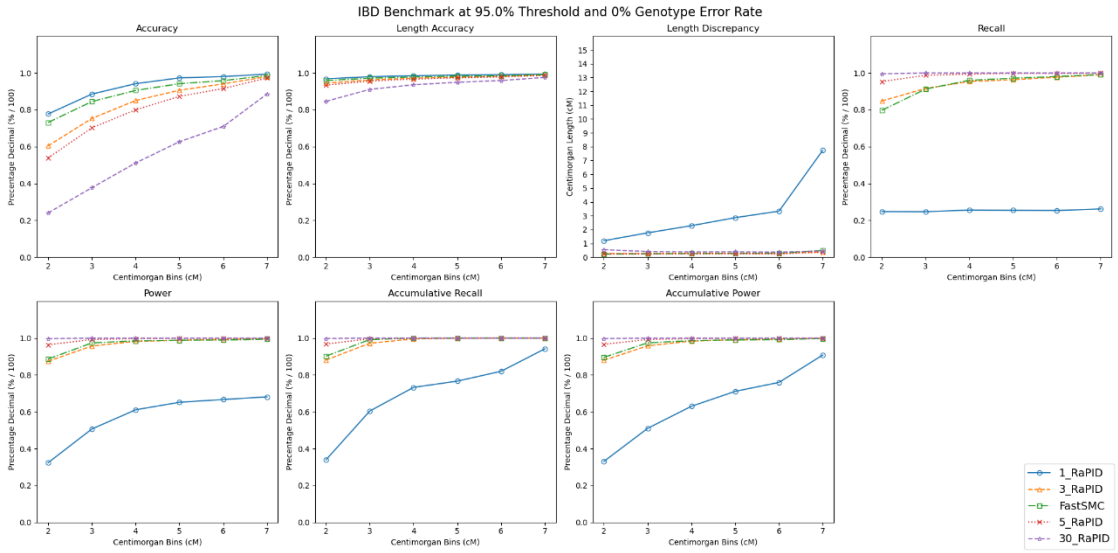

Figure 4.14: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 95% threshold and 0% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power.

At the 99% threshold (Figure 4.15), there is a significant decrease in the accuracy and recall of the tool. The tool is more accurate the smaller the window size is, however this comes at the expense of recall which experiences the opposite effect. This is because larger window sizes introduce more true positives, as well as false positives when the number of successes parameter remains constant (Naseri et al., 2019), hence effecting the recall and accuracy metrics respectively. The power metric also goes lower at the higher window sizes, while the length discrepancy is higher. There is a further decrease in the performance of the tool at the 100% threshold, represented in Figure 4.16. However, the tool still manages to achieve high recall values. In comparison to FastSMC, RaPID has better accuracy at the window size of 1, but better recall at the window sizes of 3, 5 and 30. Throughout the other metrics, RaPID's window sizes of 1, 3 and 5 perform similarly to FastSMC and to each other.

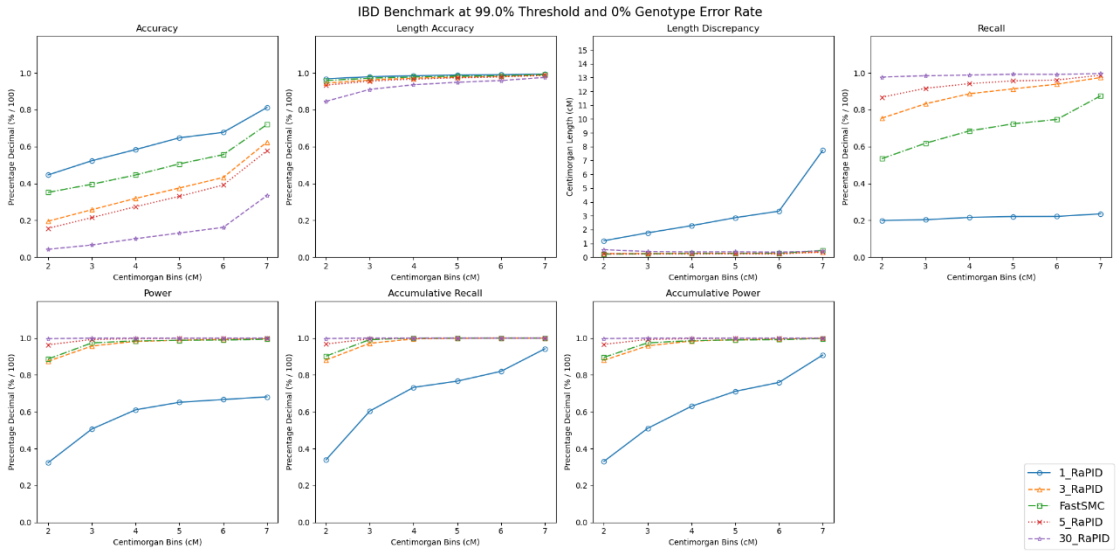

Figure 4.15: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 99% threshold and 0% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power.

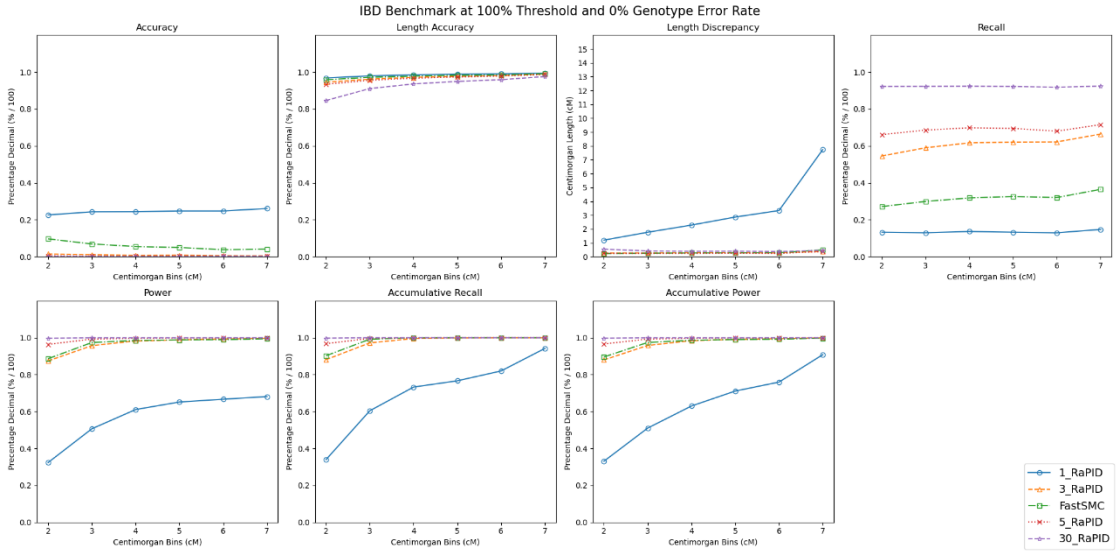

Figure 4.16: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 100% threshold and 0% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power.

4.2.2 Window Sizes at 0.01% and 0.1% Genotype Error Rates

The same window sizes of 1, 3, 5 and 30 were tested on the genotype error rates of 0.01% and 0.1%. As expected, the recall and power measures show a decrease in performance when compared to the 0% genotype error rate. This decrease is larger with 0.1% genotype error, displayed in Figures 4.17 and 4.18.

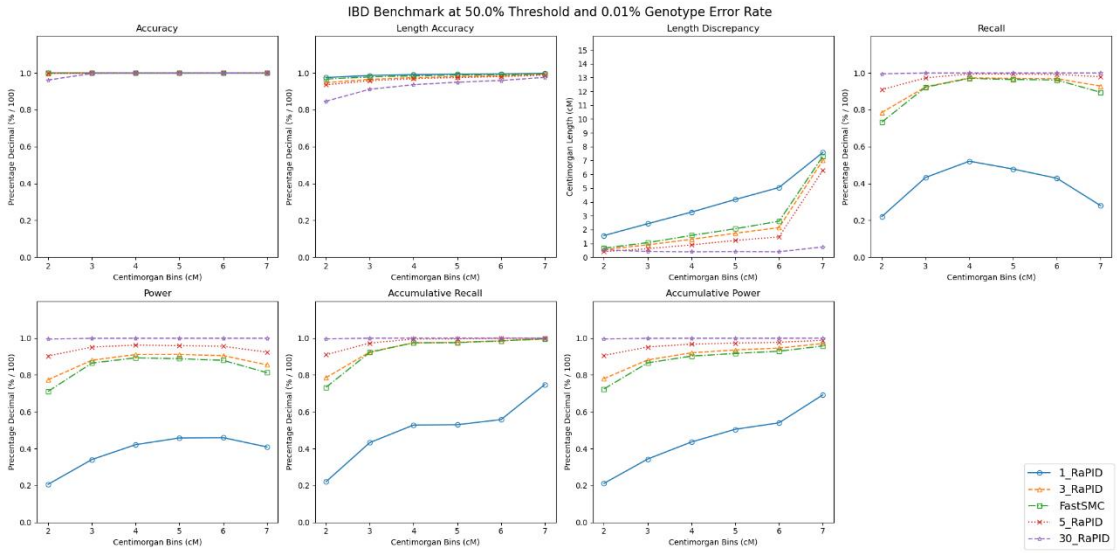

Figure 4.17: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 50% threshold and 0.01% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power.

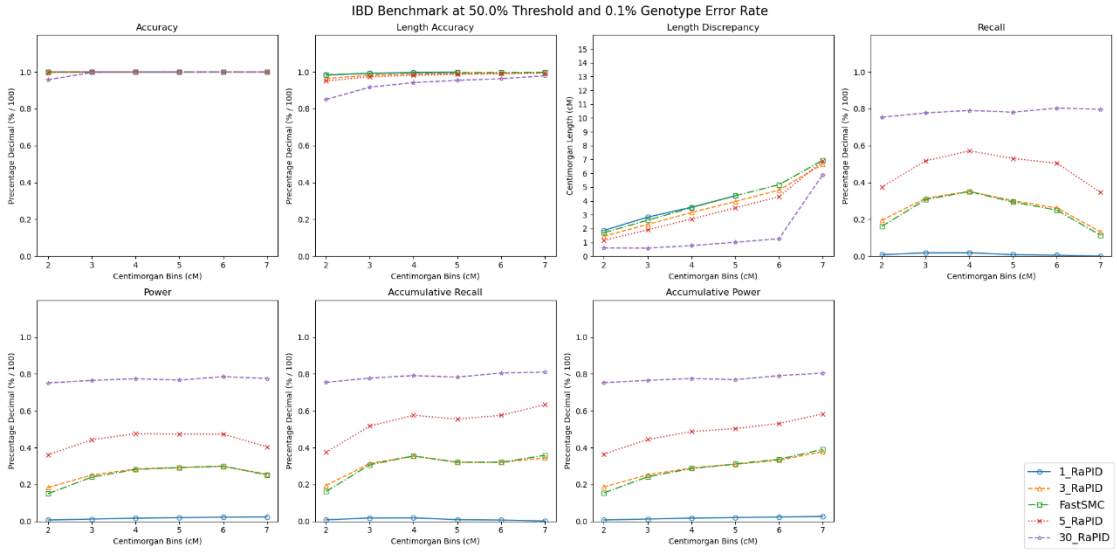

Figure 4.18: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 50% threshold and 0.1% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power.

Until the 95% threshold, RaPID's different window sizes and FastSMC show a slight decrease in performance overall when compared to the 50% threshold, as in Figures 4.19 and 4.20. This is because it becomes increasingly harder to identify larger parts of the IBD segments. RaPID's window size of 1 is unable to identify any segments larger than 5cM at the 0.1% genotype error rate. Due to the high genotype error rate, there is a higher chance that the tool encounters a genotype error. This therefore makes it more difficult for the tool

to identify the segment, especially with a window size of 1, which is the lowest amount of margin of error that RaPID allows.

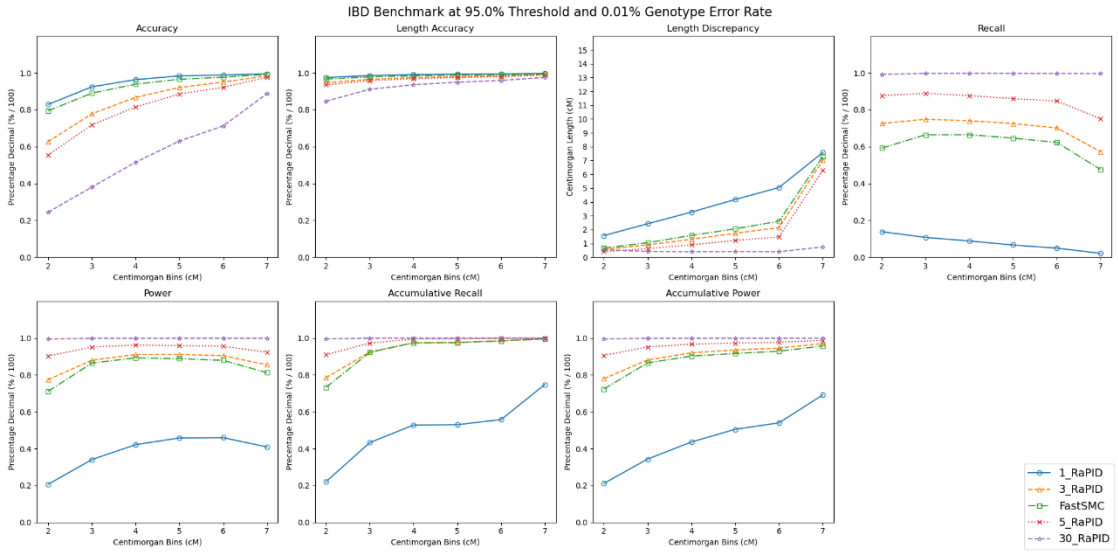

Figure 4.19: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 95% threshold and 0.01% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power. The window sizes of 3 and 5 provide a balance between accuracy, power and recall.

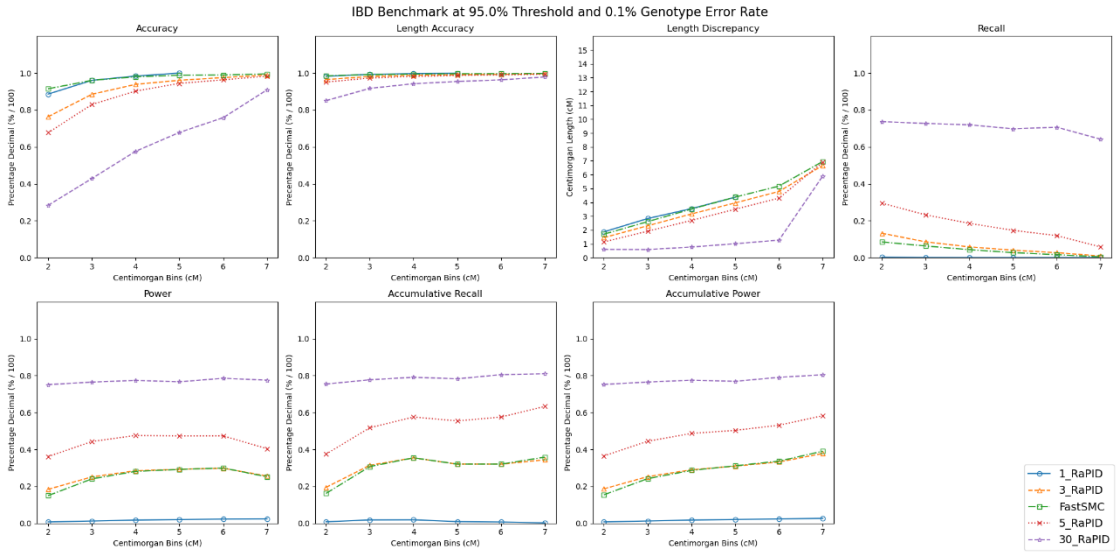

Figure 4.20: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 95% threshold and 0.1% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power. The window sizes of 3 and 5 provide a balance between accuracy, power and recall.

The 100% threshold brings a decrease in performance on the accuracy and recall parameters (Figures 4.21 and 4.22). RaPID's window size of 30 still maintains very high recall and power values, at the expense of accuracy. The

smallest window size of 1 provides the highest accuracy, accompanied with low recall and power.

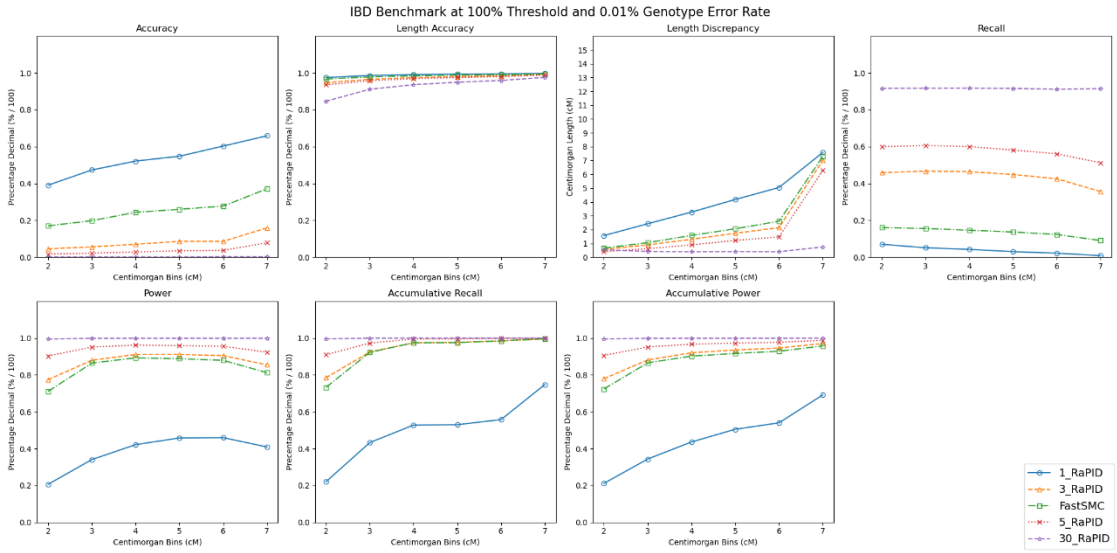

Figure 4.21: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 100% threshold and 0.01% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power. The window sizes of 3 and 5 provide a balance between accuracy, power and recall.

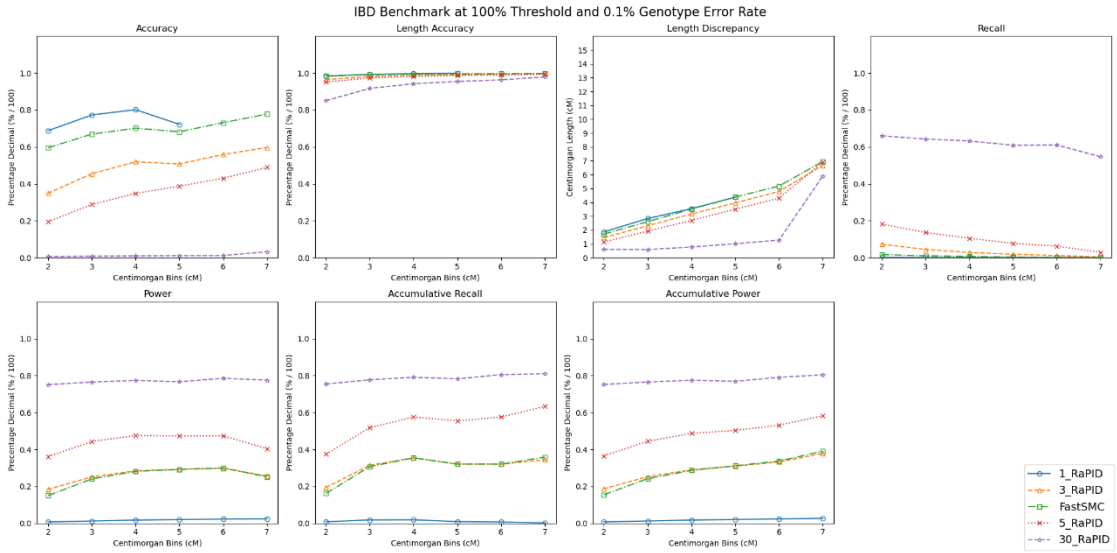

Figure 4.22: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 100% threshold and 0.1% genotype error rate. The window size of 1 has the best accuracy and length accuracy, while that of 30 has the best length discrepancy, recall, power, accumulative recall and accumulative power. The window sizes of 3 and 5 provide a balance between accuracy, power and recall.

The overall results show that as the window size of RaPID increases, so does the accuracy. However, this comes at an expense of recall and power. This is mainly seen at the higher percentage thresholds and with the introduction of

genotype error rate. The length discrepancy remains low throughout, while the length accuracy remains high. The window sizes of 3 and 5 for RaPID provide a balance between accuracy, power and recall. These window sizes were also chosen by Naseri et al. (2019) when IBD detection was performed on real UK Biobank data after benchmarking of the tool on simulated datasets was performed. With a window size of 3, RaPID performs similarly to FastSMC. RaPID's average wall-clock time of less than 4 minutes makes it the preferred tool over FastSMC's wall-clock time of more than 5 hours, therefore FastSMC can be excluded.

4.2.3 Successes Count Parameter

RaPID's window sizes of 3 and 5 that showed best accuracy and recall results were then tested further by changing the number of successes parameter. This parameter corresponds to the minimum of runs required for an IBD to be taken into consideration. The values for this parameter were tested at 2, 4 and 7, whereas the parameter for the number of runs was kept at a constant value of 10 throughout this task. The accuracy and power plots for the number of successes parameter were plotted using the Python script *rapid_sucesses.py*.

Figure 4.23 shows the performance of the tool with these parameters at 90% threshold and 0.01% and 0.1% genotype error rates. While they all show high accuracy, the window size of 5 with the number of successes set to 2 show the highest recall and power values, as well as the least length discrepancy. Naseri et al. (2019) have also validated the use of 2 as the number of successes (with the number of runs parameter set to 10) as the optimal value for this parameter to achieve the best accuracy and power, tested on a simulated dataset. These parameters should provide a balance between the generation of as many true positives as possible, while limiting the number of false positives. Therefore, they were used to perform IBD detection on the Maltese dataset.

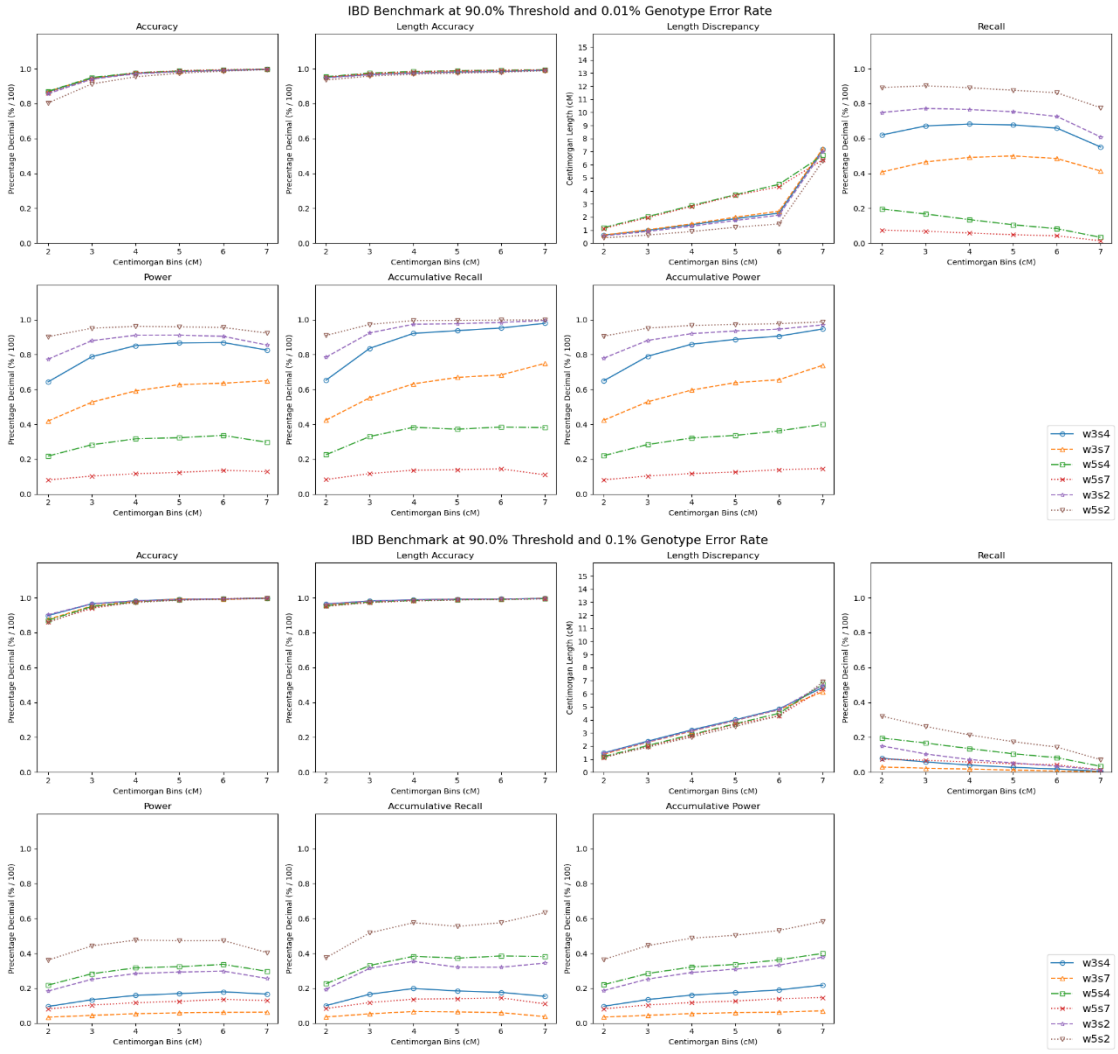

Figure 4.23: IBD Benchmark results of the RaPID's window sizes of 3 and 5 with the number of successes parameter set to 2, 4 and 7 at 0.01% and 0.1% genotype error rates. Figure legend: w = window size, s = number of successes.

4.3 Identity by Descent Detection

As per the aims and objectives, the best performing tool RaPID was used to perform IBD detection on the MAMI Maltese dataset (Attard et al., 2014). Here we will investigate whether IBD segments can be used to classify founder variants of the Maltese population. Identification of more founder variants would allow for a more specific approach in genetic testing of the Maltese population. The first set of IBD detection with RaPID was performed with the 2cM threshold, meaning that any segments shorter than 2cM would be filtered out and remain undetected. This threshold is the most commonly used by IBD detection tools

since this is capable of detecting IBD segments from the last 1,500 years (Ralph and Coop, 2013). However, this was unable to detect any IBD segments related to the compiled list of variants being investigated. This indicates that any IBD segments available for these variants are smaller than 2cM and were not detected at this threshold. Therefore, IBD detection was performed again with a threshold of 0.5cM to detect smaller segments. At this threshold, from the 15 variants analysed in this study, IBD segments were identified in 10 variants, and visualised in the form of horizontal bar plots.

IBD segments involving our compiled list of variants were visualised with the use of heatmaps, where the genotype pattern of individuals carrying the variant was compared to that of homozygous reference individuals. For a variant to be considered a founder variant, it must reside in an IBD segment which would only be found in individuals carrying the variant. These individuals would all have the same genetic framework around the variant, suggesting further that it is a founder of the population. Figures and diagrams for horizontal bar plots, heatmaps and frameworks were generated using the script *variant_search.py*.

This methodology of using IBD detection to identify founder variant is the first of its kind, meaning that direct comparative analysis of the results could not be made with other publications. Due to the ethics approval in place for the dataset being used, confirmatory analysis of founder variants using mutation age estimates was not possible as these can be directly linked to specific historical periods and the ethnicity of the population at the time. The ethics in place for this collection excluded any research related to ethnicity, therefore founder variant analysis was solely based on the use of IBD detection.

Table 4.2 shows the number of IBD segments identified by RaPID at the 2cM and 0.5cM thresholds per chromosome. There is a considerable difference between the two, with the 0.5cM threshold identifying more segments. This shows that the majority of IBD segments in the Maltese are smaller than 2cM and therefore older than 1,500 years (Ralph and Coop, 2013). The smaller the IBD segment is, the more distant the common ancestor is. This is because the segment would have gone through many recombination and mutation events,

therefore getting smaller or broken with time (Thompson, 2013). Chromosome 1 shows the largest amount of detected IBD segments while chromosome 19 has the least, possibly due to the size of the respective chromosomes.

Table 4.2: The number of IBD segments identified by RaPID at the 2cM and 0.5cM thresholds per chromosome.

Chr No. / cM Thresholds	2cM	0.5cM	Chr No. / cM Thresholds	2cM	0.5cM
Chromosome 1	781,278	6,079,254	**Chromosome 13**	36	1,044,691
Chromosome 2	292,354	3,507,306	**Chromosome 14**	343	546,735
Chromosome 3	57	458,420	**Chromosome 15**	722	2,586,947
Chromosome 4	146	1,017,505	**Chromosome 16**	75	597,614
Chromosome 5	10	591,852	**Chromosome 17**	15	310,524
Chromosome 6	6	975,312	**Chromosome 18**	127	284,309
Chromosome 7	20	880,575	**Chromosome 19**	0	88,425
Chromosome 8	0	514,533	**Chromosome 20**	72	702,274
Chromosome 9	438	523,377	**Chromosome 21**	812	2,341,644
Chromosome 10	497,678	3,534,070	**Chromosome 22**	4	683,936
Chromosome 11	11	818,988	**Chromosome X**	9,677	3,973,004
Chromosome 12	14	467,413			

4.3.1 Chromosome 1 Variant *CDCP2* p.P408RfsX46

The chromosome 1 variant *CDCP2* p.P408RfsX46 is a frameshift variant with an insertion of a T that causes an amino acid change of proline to arginine, leading to a termination codon 46 amino acids downstream. It is not related to any clinical abnormalities, but was chosen as a test variant to test the developed bioinformatics pipeline. In the Maltese population the variant has a MAF of 0.008, whereas in other populations the MAF of this variant is almost equal to 0, hence why it was chosen to test its founder status. Of the 14 heterozygote individuals present in our dataset, five show the presence of an IBD segment detected by RaPID which goes through the position of the variant. These are accompanied by another four homozygous reference individuals and represented as pairwise IBD segments in Figure 4.24. The positions (highlighted with an orange vertical line) going through the identified IBD segments are common in all of the IBD segments and are used to visualise the genotype of

individuals around the area of the variant of interest in the form of a heatmap (Figure 4.25).

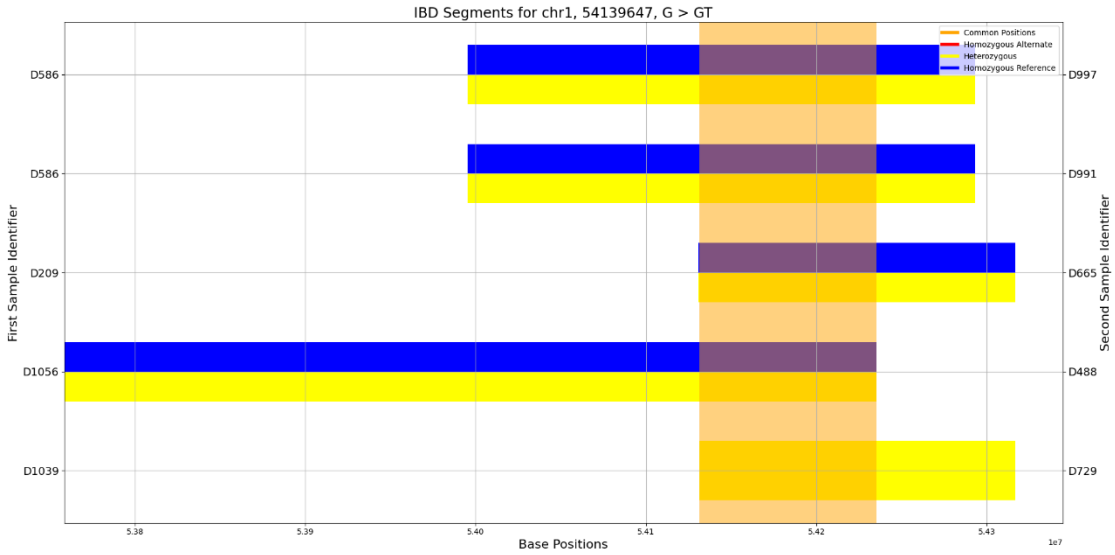

Figure 4.24: Horizontal bar plots showing the detected IBD segments for the chromosome 1 position 54,139,647. This includes 5 heterozygous individuals for the variant CDCP2 p.P408RfsX46.

Figure 4.25 represents a heatmap around the area of the variant which lies within the common position range of the IBD segment. This includes all of the individuals carrying the variant (in this case only heterozygous individuals) and a random 50 homozygous reference individuals chosen by equal probability sampling. The x-axis consists of the basepair positions whereas the y-axis consists of the sample identifiers. This heatmap uses the genotypes of individuals to identify whether the variant of interest forms part of the detected IBD segment or opposite to it on the other allele.

The IBD segment identified by RaPID lies on the opposite allele that does not carry the variant. The variant *CDCP2* p.P408RfsX46 is most likely not a founder variant of the Maltese population, since it does not form part of the IBD segment presented. However, being chosen as a test variant, this shows that any variant can be investigated for its founder status with the developed bioinformatics pipeline, as long as the variant is present in the dataset and inserted in the tool in the correct format.

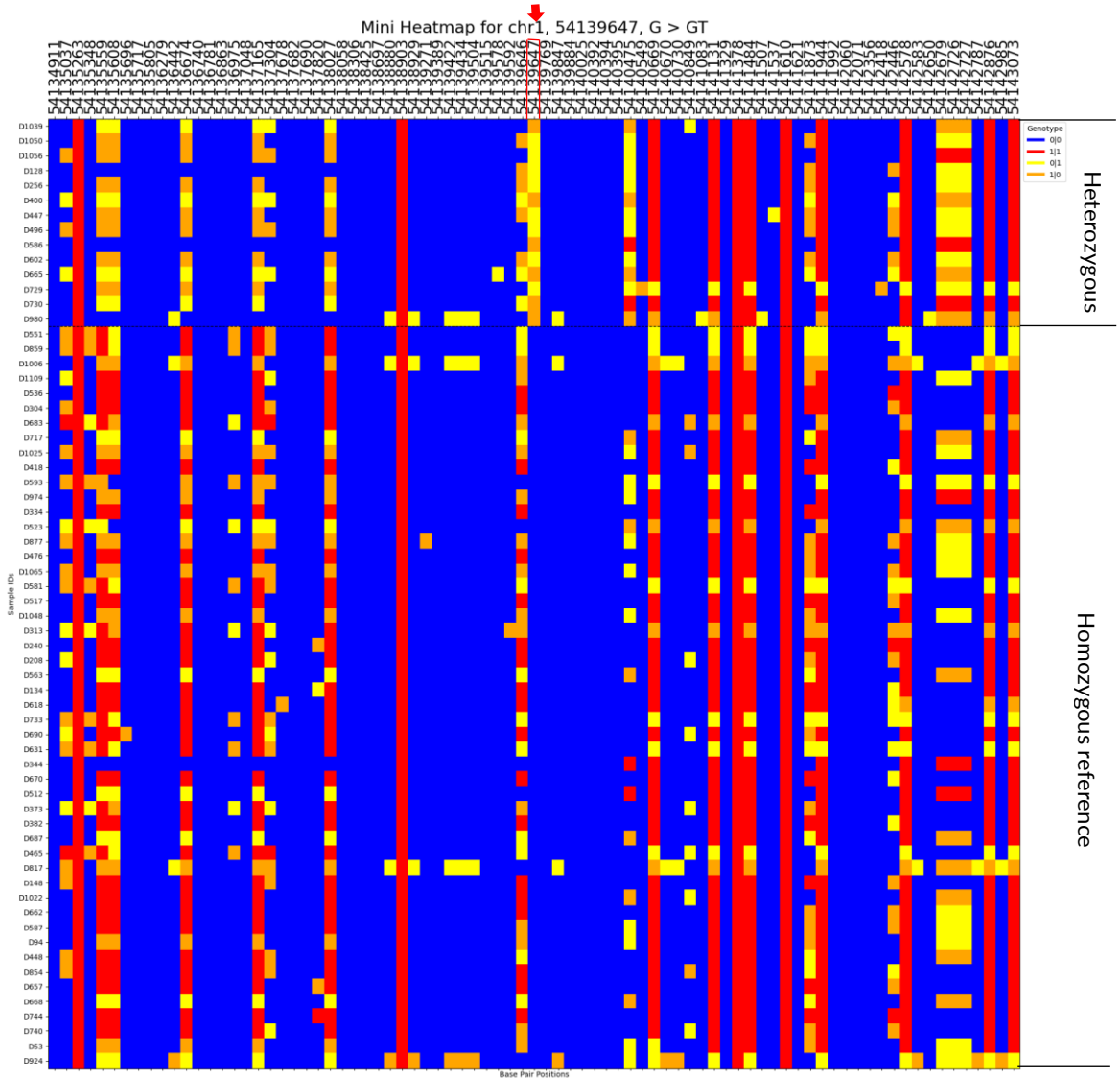

Figure 4.25: Heatmap showing the common IBD segment positions for the chromosome 1 variant CDCP2 p.P408RfsX46. This includes 14 heterozygous and a random 50 homozygous reference individuals. The detected IBD segment resides on the opposing allele that does not carry the variant, therefore this variant is not a founder variant. The variant is indicated by a red arrow.

The subsequent variants that will be described were all chosen because of their relevance for diseases present in the Maltese population and ongoing interest from other research projects.

4.3.2 Chromosome 1 Variants *KISS1* p.X139fs, *KISS1* p.P81R and *KISS1* p.Q36R

The three chromosome 1 variants *KISS1* p.X139fs, *KISS1* p.P81R and *KISS1* p.Q36R are found only a few hundred basepair positions apart. Being close to each other, there is a high probability that these variants are inherited together

and thus forming part of the same IBD segment. *KISS1* is involved in the regulation of gonadotropin-releasing hormone (GnRH) release and in the regulation of the hypothalamic-pituitary-gonadal axis. Variants in the gene influence the levels of GnRH, and in turn the levels of reproductive hormones such as FSH and LH. This can lead to several reproductive-related diseases such as IHH and infertility (Zhu et al., 2022).

The first variant p.X139fs occurs due to a deletion of a T at the terminal codon of the gene and results in a stop-loss associated with high anti-Mullerian hormone, in both homozygote alternates and heterozygote individuals (Trevisan et al., 2020). The presence of the variant also results in an increase in ovarian progesterone and prolactin in females, which are also regulated by *KISS1*. High levels of these hormones are known to inhibit the production of FSH and LH from the pituitary gland and therefore may cause infertility. In fact, Trevisan et al. found the variant to be related to a decrease in the number of oocytes produced in the ovaries and a decrease in the number of successful pregnancies.

The second variant p.P81R with a SNP of G to C, like many other *KISS1* variants, is characterized by a deregulation of kisspeptin, which is a protein product of the gene. Its main role is to bind to GnRH receptors in order to regulate the release of FSH and LH, and disruption of this process can result in IHH and the development of polycystic ovarian syndrome (PCOS) (Stephen et al., 2024). The latter is a common disorder in women of reproductive age, accounting for a prevalence of around 6-12% (Meng et al., 2023; Wojciechowski et al., 2012). The variant has also been linked with an unexplained recurrence of pregnancy loss, and is suspected to be a risk factor (Meng et al., 2023).

The third chromosome 1 variant *KISS1* p.Q36R is a missense variant characterized by a SNP of a T to C nucleotide, and is associated with reproductive hormones. Women being heterozygous for this variant have shown low levels of LH and produced more oocytes when compared to the wild-type women (Martins Trevisan et al., 2020). It has also been confirmed through multiple studies that this variant does not have an association with PCOS (Daghestani et al., 2020; Farsimadan et al., 2021; Krstevska-Konstantinova et al., 2014).

Having a high MAF of 0.255 and 0.298 in the Maltese respectively, the *KISS1* p.X139fs and p.P81R variants are slightly overrepresented when compared to the MAFs of 0.214 and 0.252 in non-Finnish Europeans (NFE). The *KISS1* p.Q36R however has a MAF of 0.063, which is remarkably lower than the other two variants. This suggests that the *KISS1* p.X139fs and p.P81R variants may be in LD and possibly even part of the same IBD, whereas the *KISS1* p.Q36R is not.

Figures 4.26 to 4.28 show the pairwise IBD segments that were identified by RaPID at the basepair positions of the variants. The common positions going through the identified IBD segments are the same in the *KISS1* p.X139fs and p.P81R variants, ranging from basepair position 204,190,416 to position 204,193,102, with both of variants falling within this range. For the variant at position 204,190,483 (p.X139fs), 305 heterozygous and 54 homozygous alternate cases were present in the dataset, while the variant at position 204,190,659 (p.P81R) had 335 heterozygous cases and 78 homozygous alternate cases. Of these, RaPID detected an IBD segment in 115 heterozygotes and 20 homozygous alternate cases (38% of individuals with the variant) and 131 heterozygotes and 24 homozygous alternate cases (28% of individuals with the variant) respectively for each variant. A ROH containing both variants is identified in sample D202, spanning more than 1,500,000 basepairs. For the p.Q36R variant, IBD segments were detected in 19 heterozygous and one homozygous alternate individual at position 204,190,794. This comprises of only 18% of the individuals with the variant in the dataset, consisting of 106 heterozygous & six homozygous alternate individuals. A ROH containing the variant was detected, spanning just over 500,000 basepairs.

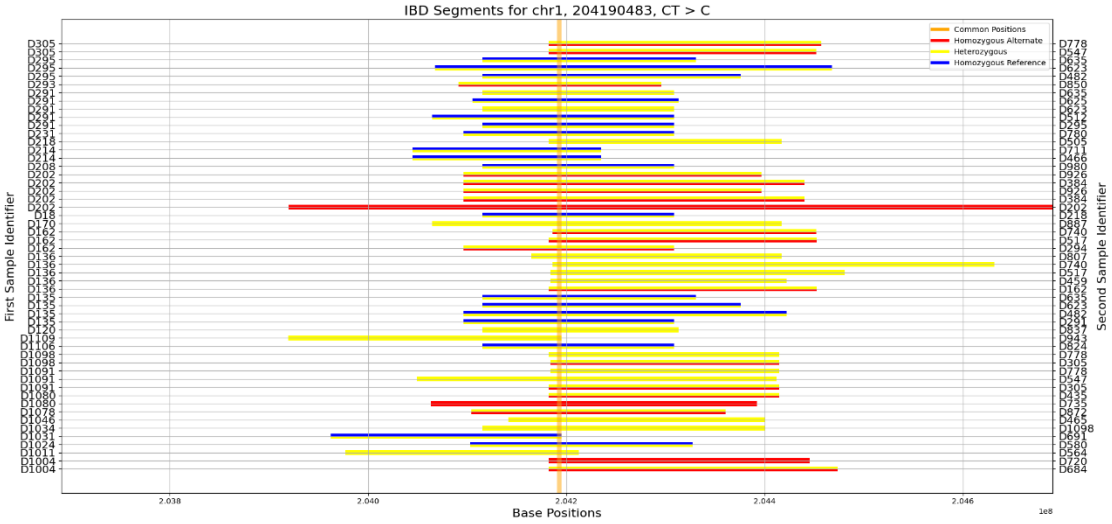

Figure 4.26: A representative horizontal bar plot showing the detected IBD segments for chromosome 1 position 204,190,483. This was extracted from a total of 115 heterozygotes and 20 homozygous alternate individuals for the variant KISS1 p.X139fs. A ROH is identified in sample D202.

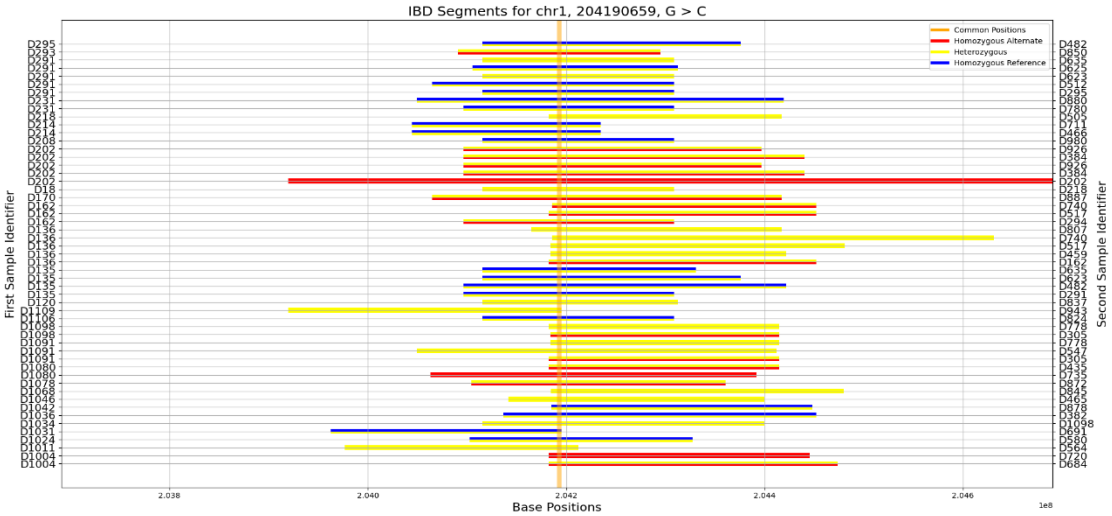

Figure 4.27: A representative horizontal bar plot showing the detected IBD segments for chromosome 1 position 204,190,659. This was extracted from a total of 131 heterozygotes and 24 homozygous alternate individuals for the variant KISS1 p.P81R. A ROH is identified in sample D202.

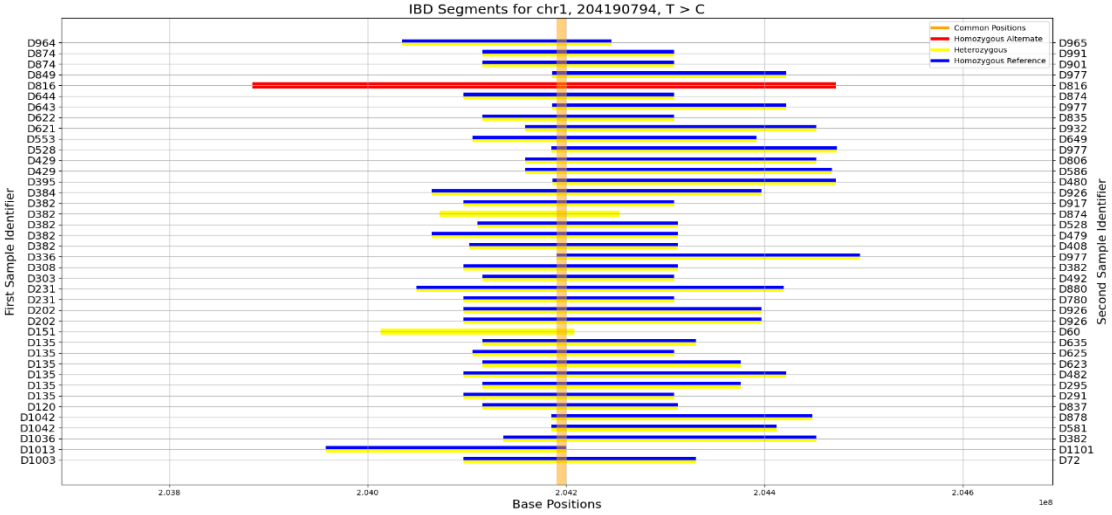

Figure 4.28: Horizontal bar plot showing the detected IBD segments for the chromosome 1 position 204,190,794. This includes 19 heterozygous and 1 homozygous alternate individuals for the variant KISS1 p.Q36R. A ROH is identified in sample D818.

Figures 4.29 and 4.30 show the heatmaps for the *KISS1* variants p.X139fs and p.P81R for the common position range. This includes all 844 individuals which are divided into three representative groups of homozygous reference, heterozygous and homozygous alternate individuals. There is a clear distinction between the three groups, showing the presence of an IBD segment carrying both of the variants in almost all of the individuals that have them. This is also absent from the homozygous reference cases. This suggests that the heterozygous and homozygous alternate individuals share a common ancestor and that both of the variants are most likely founder variants. The figures also show that the *KISS1* p.X139fs is always in conjunction with *KISS1* p.P81R, but some heterozygous individuals only carry the latter variant, without the former variant. This indicates that the *KISS1* p.P81R variant is the older variant of the two, and that the variants originated from two different common ancestors. Both heatmaps also show that the *KISS1* p.Q36R variant (position 204,190,794) does not form part of this IBD segment.

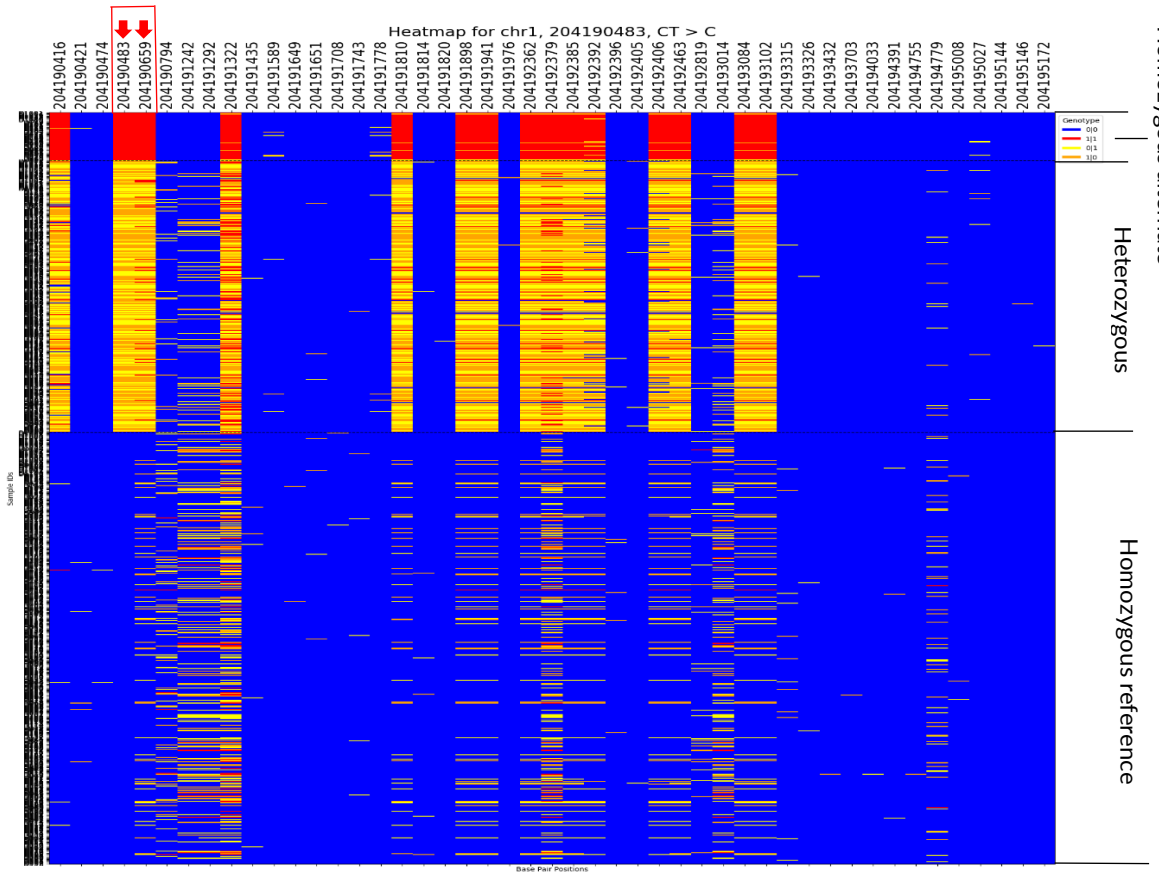

Figure 4.29: Heatmap showing the common IBD segment positions for KISS1 p.X139fs, indicating an IBD segment in the individuals with the variant. The variants are indicated by red arrows.

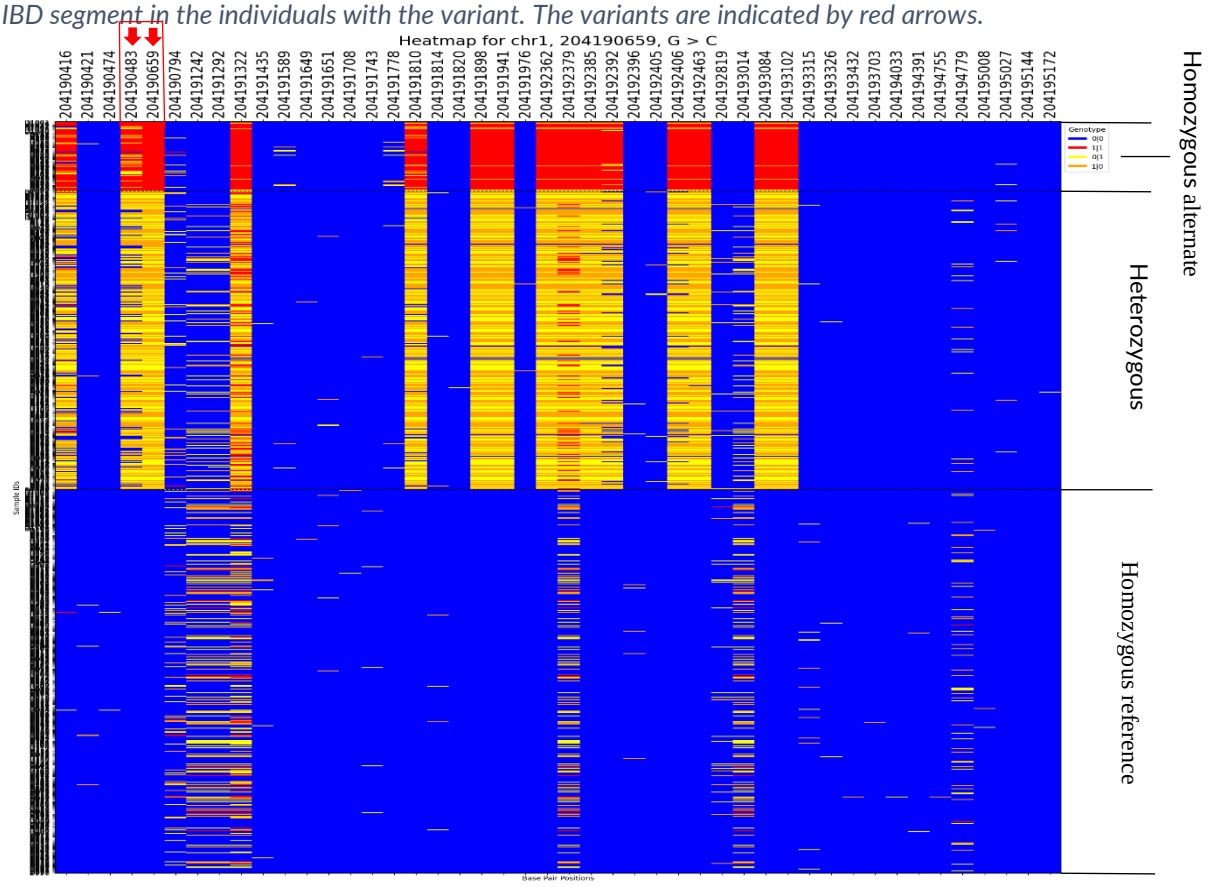

Figure 4.30: Heatmap showing the common IBD segment positions for KISS1 p.P81R, indicating an IBD segment in the individuals with the variant. The variants are indicated by red arrows.

The genetic framework for this common position range of the IBD segment is represented in Figure 4.31. This comprises of the alleles that are on the same chromosome as the variant, and in this case the detected IBD segment. This is common for both the *KISS1* p.X139fs and p.P81R variants, but excludes the p.Q36R which does not form part of the IBD segment.

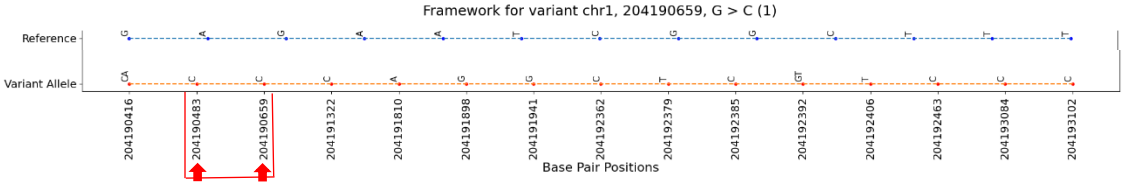

Figure 4.31: The IBD segment variant framework for the chromosome 1 variants KISS1 p.X139fs and KISS1 p.P81R. The variants are indicated by red arrows.

The heatmap in Figure 4.32 showcases the common positions of the IBD segments for *KISS1* p.Q36R. It shows that the variant does not form part of the IBD segment which was identified by RaPID, containing the *KISS1* variants p.X139fs and p.P81R. The variant is always present on the allele opposite to the one carrying the IBD segment, and this segment is present in many homozygous reference cases. This indicates that the variant does not form part of the IBD segment related to the previously mentioned *KISS1* variants.

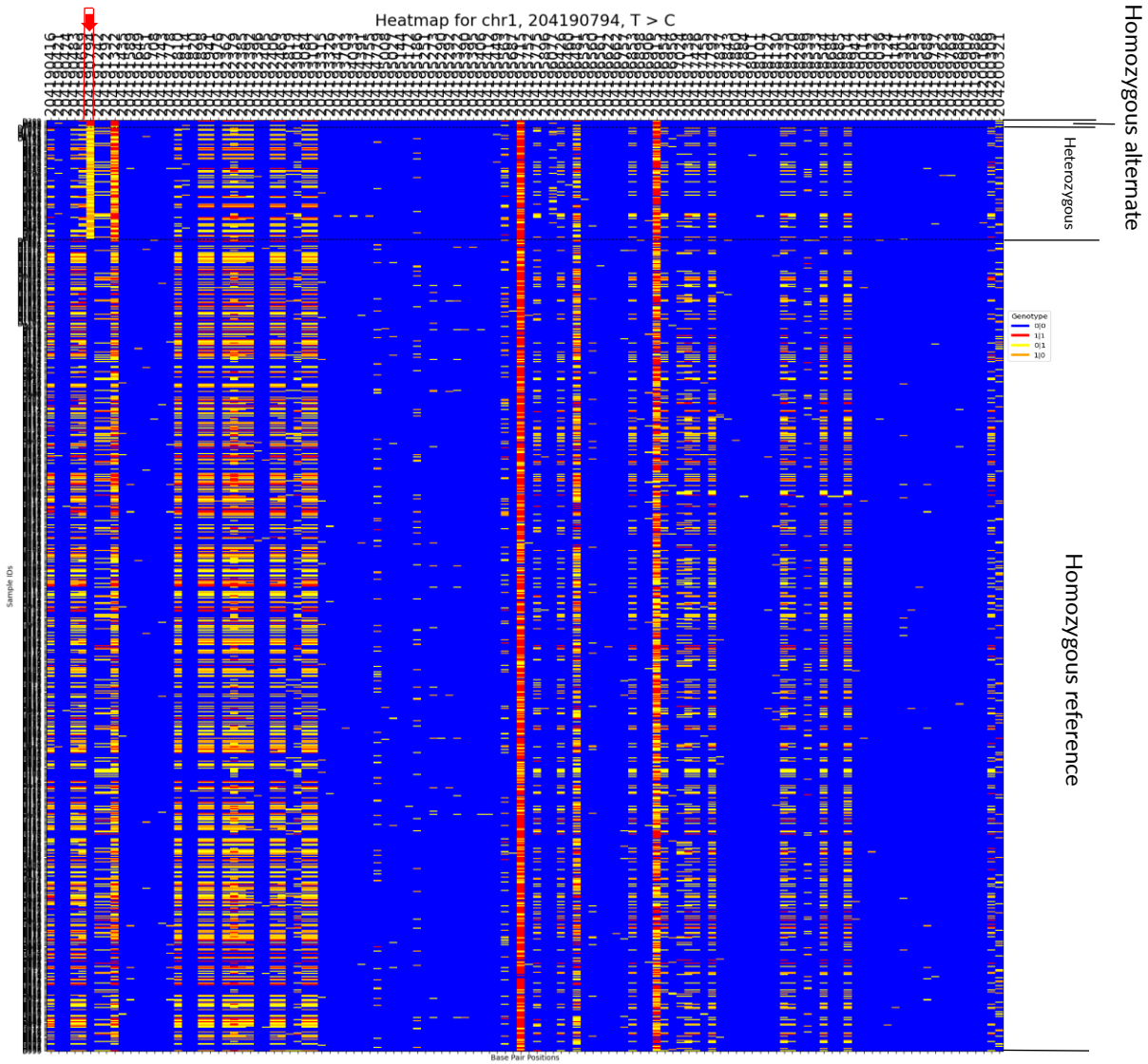

Figure 4.32: Heatmap showing the common IBD segment positions for the variant KISS1 p.Q36R, indicating that the variant is not part of the detected IBD segment as it is found on the opposite allele. The variant is indicated by a red arrow.

The genetic framework in Figure 4.33 for this range of positions is a confirmation of the above results, showing that the p.Q36R variant has no link with the other two *KISS1* variants that have a completely different framework.

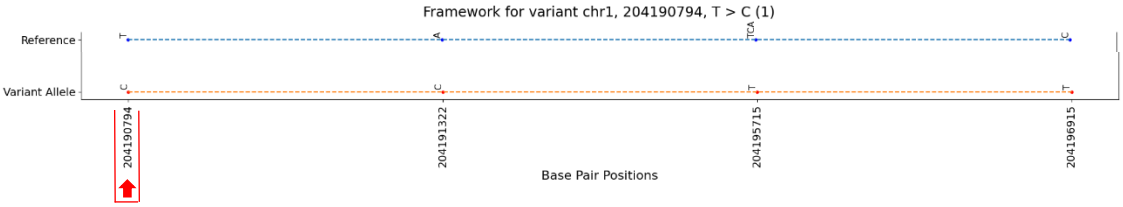

Figure 4.33: The variant framework for individuals that carry the chromosome 1 variant KISS1 p.Q36R. This is different than the framework presented for the KISS1 variants p.X139fs and p.P81R and thus does no share their IBD segment. The variant is indicated by a red arrow.

Even though no IBD segments for the *KISS1* p.X139fs and p.P81R variants were identified by RaPID at the 2cM threshold, this still detected a ROH in

sample D202 showcasing both of the variants. This is the same segment identified at the 0.5cM threshold (Figures 4.26 and 4.27). It ranges from position 203,921,160 to position 205,068,226, totalling more than 1,700,000 basepairs. This also does not include the *KISS1* p.Q36R variant at position 204,190,794.

Confirming the presence of an IBD segment in approximately 43% of the entire dataset, containing both *KISS1* variants, as well as obtaining a genetic framework for it, there is strong evidence in favour of these variants having a founder status in the Maltese population. Krstevska-Konstantinova et al. (2014) found both of these variants together in 28 females diagnosed with idiopathic central precocious puberty (ICPP), which is characterised by puberty before the age of eight years, as well as in the control groups which consisted of patients with normal puberty development, suggesting that they are not associated with ICPP. So far, no other publication has tried to link these two variants together.

4.3.3 Chromosome 2 Variant *SPR* c.596-2A>G

The chromosome 2 *SPR* c.596-2A>G variant at basepair position 72,891,345 is a splice acceptor variant. It was identified in seven Maltese children who showed symptoms of cognitive impairment and early motor delay. The single nucleotide change of the second nucleotide in the exon-intron junction causes aberrant splicing resulting in a lack of Sepiapterin reductase production, which is an enzyme involved in the production of BH_4 (Neville et al., 2005). Decrease in BH_4 causes hyperphenylalaninaemia with neurotransmitter deficiency, which if left untreated can lead to brain and nerve damage (Farrugia et al., 2007). Being a pathogenic variant mainly recorded in the Maltese, it is possible that this variant is a founder variant.

Out of the 844 individuals present in the dataset, 18 individuals were heterozygous for the variant. No homozygous alternate samples were present. Figure 4.34 showcases the IBD segments that were detected for this variant, involving five heterozygous individuals that each share an IBD segment with five homozygous reference individuals. This gives an indication that the detected IBD

segment is not on the allele that carries the variant, but on the allele with the reference sequence. Between them, the identified individuals share the majority of the IBD segment which ranges approximately 270,000 basepairs.

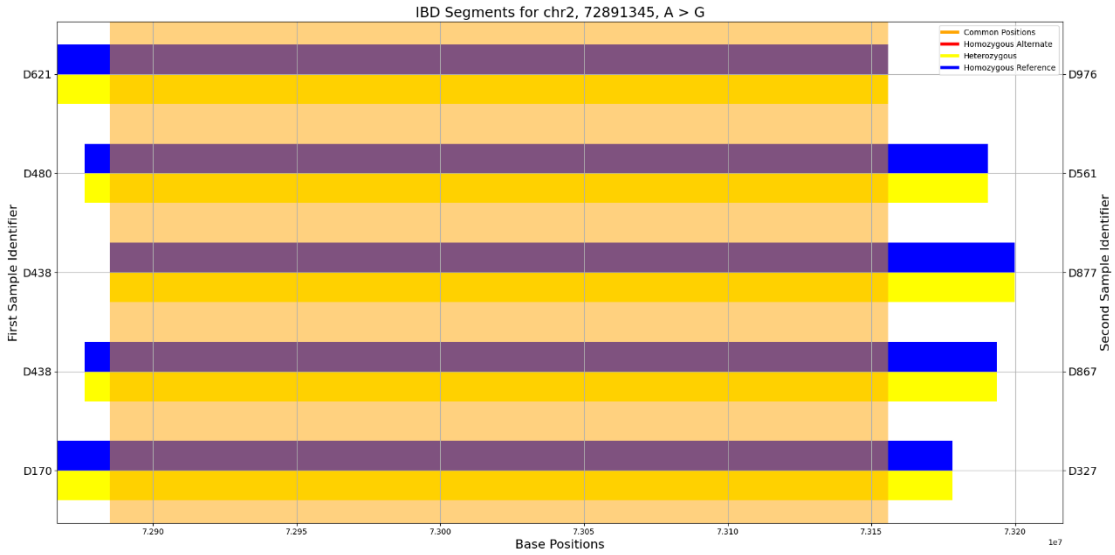

Figure 4.34: Horizontal bar plots showing the detected IBD segments for the chromosome 2 position 72,891,345. This includes 9 heterozygous individuals for the variant SPR c.596-2A>G.

Using the common range of positions for the detected IBD segments, which covers almost the entire segment, a heatmap for all the heterozygous and a random 50 homozygous reference individuals from the dataset was plotted, focusing on the area around the variant (Figure 4.35). The heatmap confirms that an IBD is present on the allele that does not carry the variant. It is inconclusive as to whether an IBD segment exists on the allele of the variant, and based on these results, we cannot suggest a founder status for it. Being a pathogenic variant which has mainly been recorded in the Maltese population, it is possible that the variant is very old and forms part of a very small IBD segment which is undetectable by the tool.

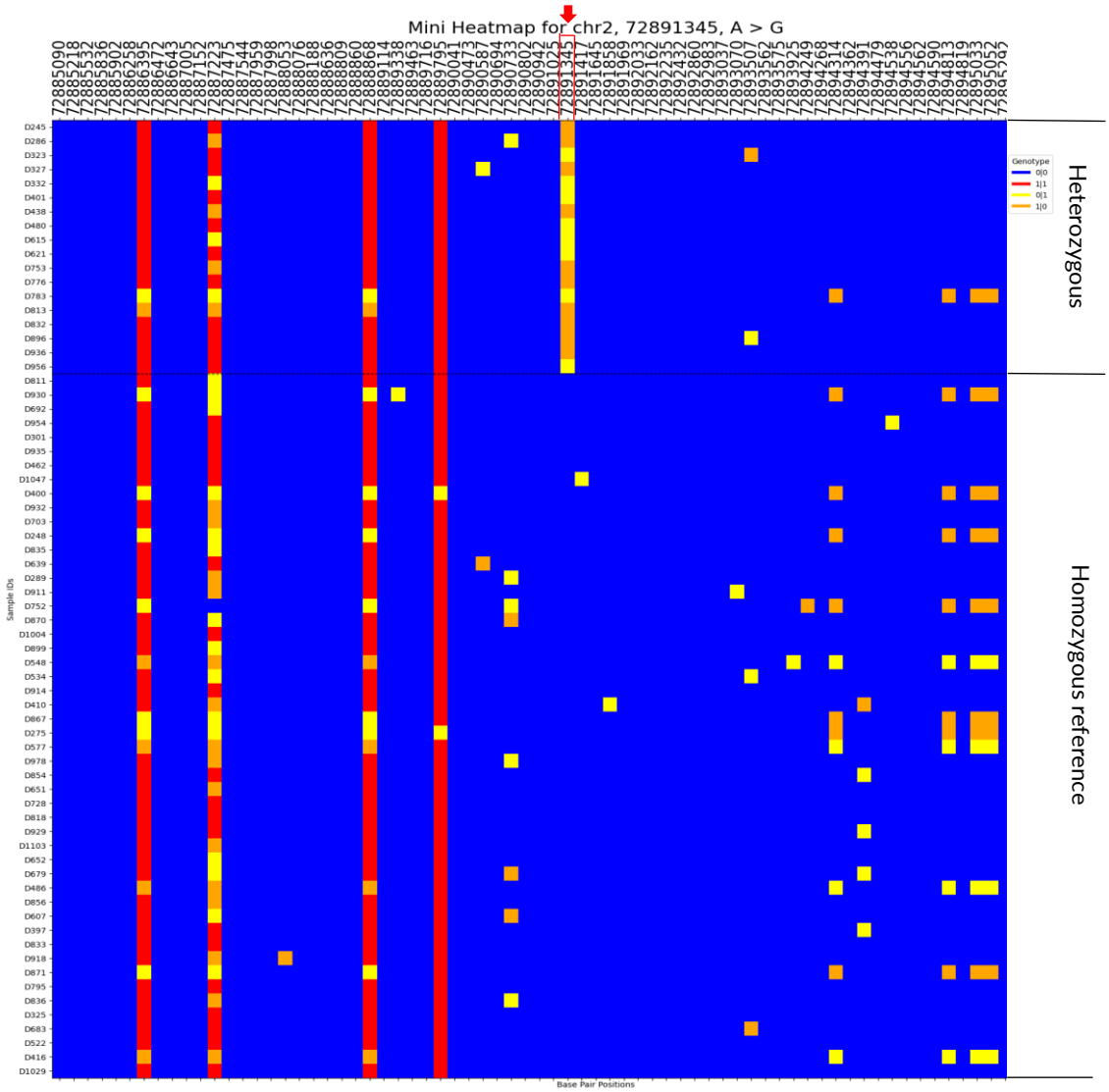

Figure 4.35: Heatmap showing the common IBD segment positions for the chromosome 2 variant SPR c.596-2A>G. This includes 18 heterozygous individuals and a random 50 homozygous reference individuals. An IBD segment is present on the opposite allele that does not carry the variant, but otherwise the results of this heatmap are inconclusive. The variant is indicated by a red arrow.

4.3.4 Chromosome 4 Variant *GNRHR* p.Q106R

The variant on chromosome 4 *GNRHR* p.Q106R is a genomic SNP of a T to C. This gene encodes the receptor for gonadotropin-releasing hormone (GnRH), with the latter inducing the pituitary gland to produce FSH and LH. A disruption in the GnRH receptors leads to IHH, which is characterized by the partial or total lack of development during puberty. This in turn causes issues in reproduction and sexual maturation (Chevrier et al., 2011). The SNP of a T to a C nucleotide leads to a substitution of glutamine to arginine (p.Q106R) on the first extracellular hydrophobic loop of the G protein-coupled receptor (Jardón-

Valadez et al., 2008). This causes a change in the receptor's shape which in turn reduces ligand binding and receptor activation, therefore causing partial loss-of-function (de Roux, 2006).

GNRHR p.Q106R been reported as being the most commonly identified pathogenic variant in *GNRHR*, and can be found in many different populations across Europe, America and South Asia. Axiak et al. (2023) analysed high throughput sequencing data of 146 Maltese individuals, of whom 17 were part of an IHH cohort. Of these, eight heterozygotes for this *GNRHR* variant were identified, four of which forming part of the patient cohort. A local study was also performed on 493 Maltese cord blood DNA, where 25 heterozygous and two homozygous alternative unrelated individuals were identified. This resulted in a MAF of 0.029, which is higher than any other population. The highest reported MAF outside of Malta is that of 0.005 in southern Europe (Karczewski et al., 2020). This motivated Axiak et al. (2023) to study 978 individuals from the MAMI study (Attard et al., 2014), where 43 heterozygotes (26 men and 17 women) were found. This translated to a MAF of 0.033, suggested that this variant is a probable founder variant due to its high overrepresentation in the Maltese. The majority of the individuals did not experience any distinctive characteristics from homozygous wild-type individuals. The heterozygosity of the variant did not seem to have an effect on the levels of gonadotropins and sex steroid hormones, therefore their fertility remained unaffected. However, studies have shown that homozygosity for *GNRHR* partial loss-of-function variants have an association with late menarche and puberty delay (Gianetti et al., 2012; Howard, 2019), highlighting the importance of identifying such cases.

Figure 4.36 showcases the IBD segments that were detected by RaPID at the variant's location. There are 41 heterozygous and one homozygous alternate individuals for the variant in our dataset. Of these, four heterozygotes and the one homozygous alternate individuals were identified to have an IBD segment, together with one homozygous reference individual. Sample D878 contains a ROH which spans at least 750,000 basepairs.

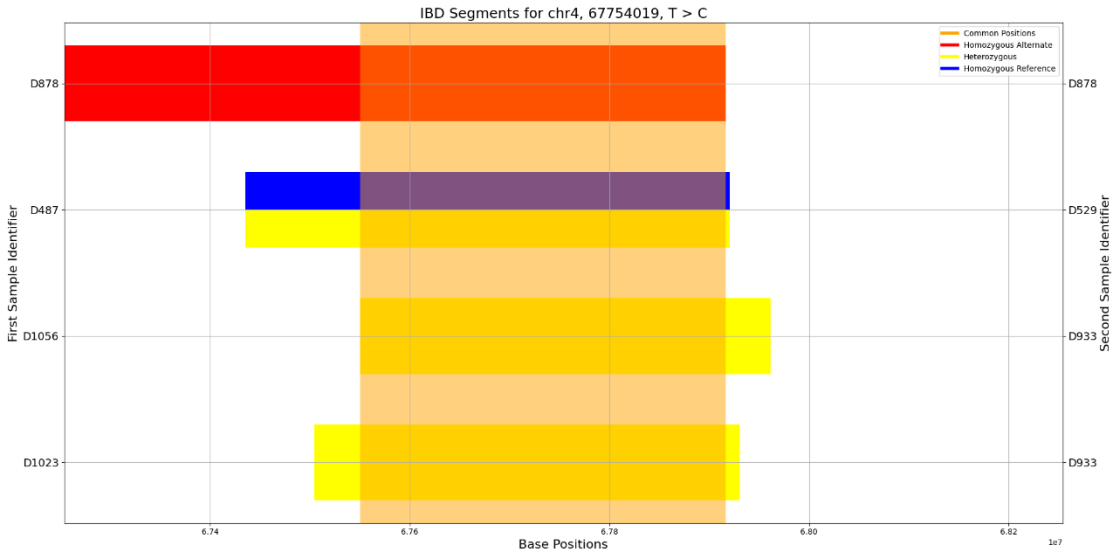

Figure 4.36: Horizontal bar plots showing the detected IBD segments for the chromosome 4 position 67,754,019. This includes 4 heterozygous and 1 homozygous alternate individuals for the variant GNRHR p.Q106R.A ROH was identified in sample D878.

A heatmap to showcase all individuals carrying the variant and a random 50 homozygous reference individuals from the dataset was plotted, focusing around the area of the variant within the common positions of the IBD segments (Figure 4.37). This shows the presence of two genetic frameworks for the variant. The larger framework was most probably detected by RaPID over the smaller framework due to its small size (smaller than 0.5cM threshold used).

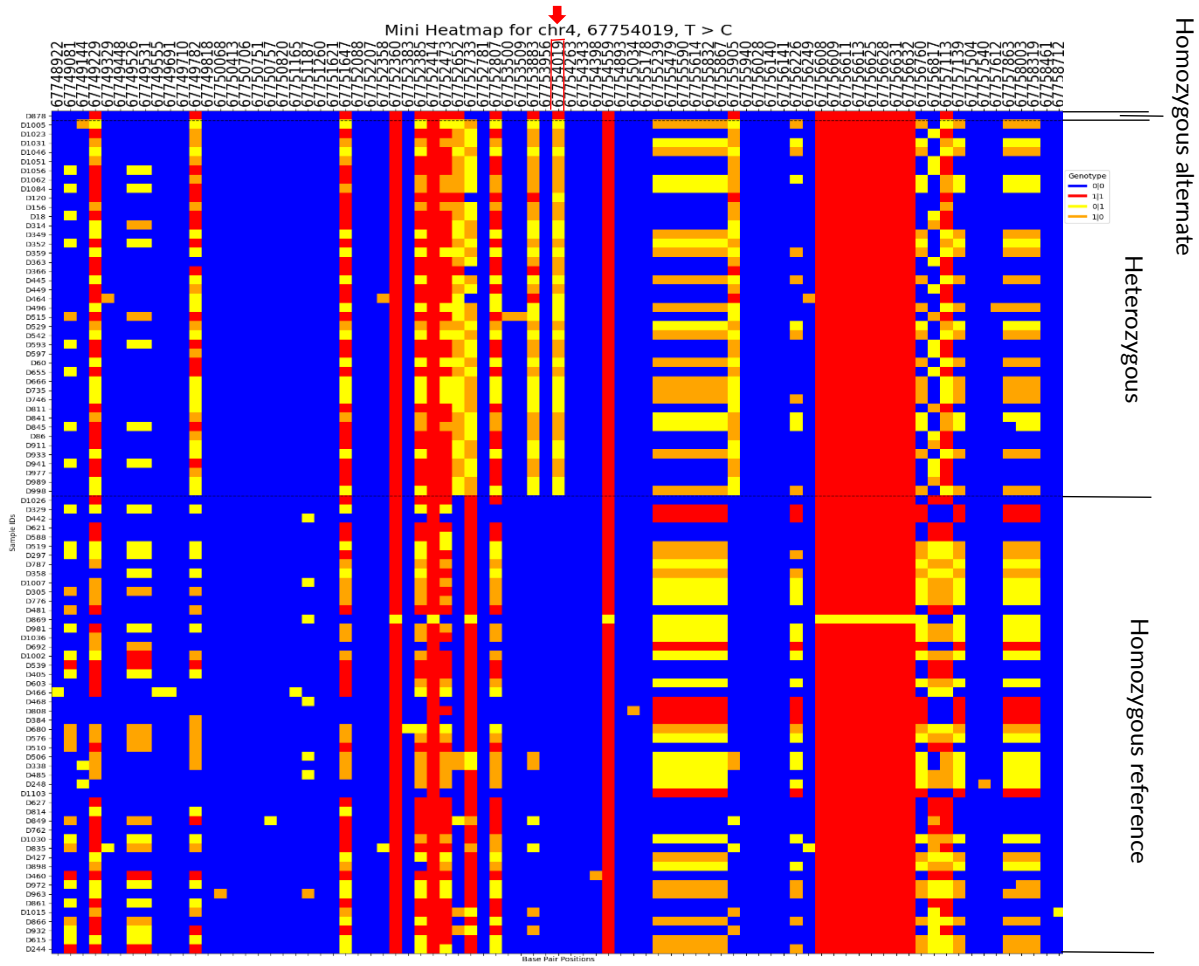

Figure 4.37: Heatmap showing the common IBD segment positions for the chromosome 4 variant GNRHR p.Q106R. This includes 1 homozygous alternate, 41 heterozygous and a random 50 homozygous reference individuals. This shows two IBD segments, the smaller of which residing on the allele that carries the variant. The variant is indicated by a red arrow.

Figure 4.38 showcases the variant framework, which is common in all of the individuals that are carrying the variant and unique to the variant's allele. Being overrepresented in the population, as well as confirming its presence as part of a genetic framework, the *GNRHR* p.Q106R variant is most likely a founder variant of the Maltese.

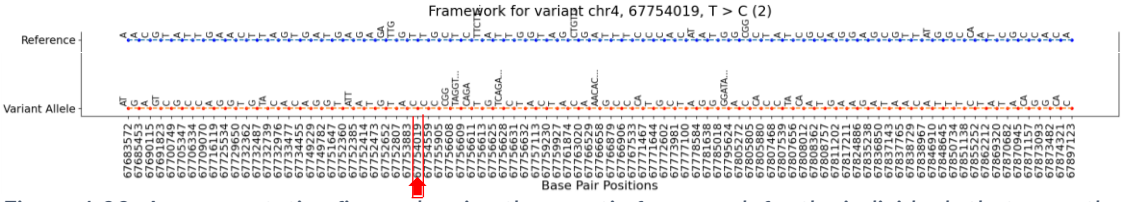

Figure 4.38: A representative figure showing the genetic framework for the individuals that carry the chromosome 4 variant GNRHR p.Q106R. The variant is indicated by a red arrow. Abbreviated indels: insertion of AGGTATGG at 67,756,609, deletion of TCTTTATA at 67,756,613, insertion of CAGAAA at 67,756,628, insertion of CAGA at 67,766,658, insertion of GATAGATA at67,795,624.

4.3.5 Chromosome 4 variant *TACR3* p.K286R

The chromosome 4 variant *TACR3* p.K286R is found at position 103,656,225 as a single nucleotide missense variant of a T to a C. The *TACR3* gene is one of three genes which encode for tachykinin receptors. Tachykinin neurokinin 3 binds to the *TACR3* receptors and is involved in the control of reproductive neuroendocrine functions (Lasaga and Debeljuk, 2011). The SNP at this position causes an amino acid change of a lysine to arginine, and has been associated with central precocious puberty, IHH and constitutional delay of growth and puberty, with a frequency of 0.9%, 1.4% and 2% of patients respectively (Tusset et al., 2012).

The variant has a MAF of 0.008 in the Maltese population, higher than the MAF of 0.001 in NFE. Fifteen heterozygous individuals were present in the dataset, with no homozygous alternate cases. One IBD segment was identified that goes through the variant position (Figure 4.39). This includes one heterozygous individual and one homozygous reference individual for the variant.

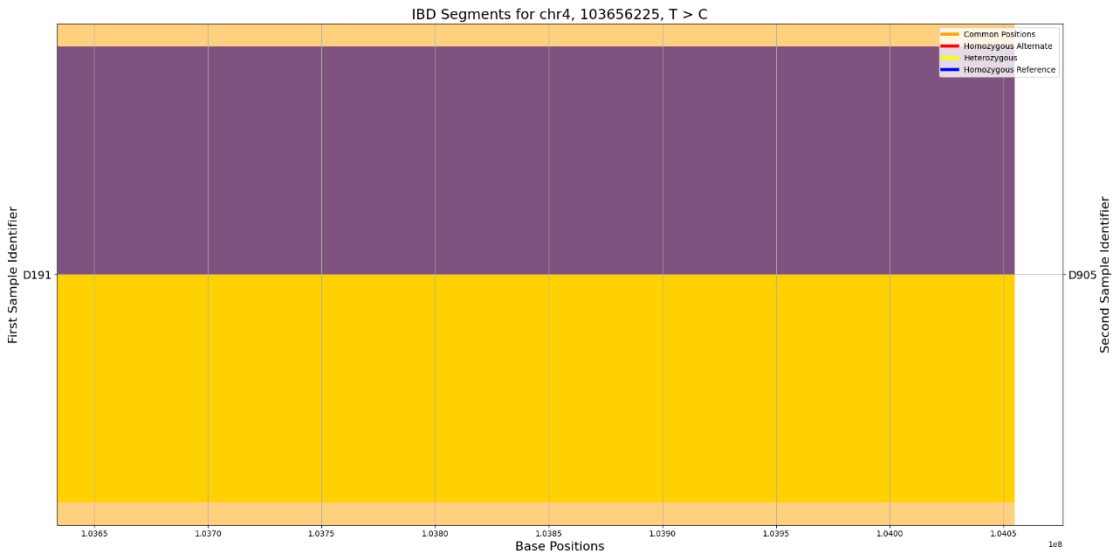

Figure 4.39: Horizontal bar plot showing the detected IBD segments for the chromosome 4 position 67,754,019. This includes 1 heterozygous individual for the variant TACR3 p.K286R, paired with a homozygous reference individual.

Figure 4.40 shows the heatmap around the area of the variant which lies within the common position range of the IBD. This includes all heterozygous individuals and a random 50 homozygous reference individuals. The heatmap

shows a genetic framework on the allele that carries the variant, which was not detected as an IBD segment by RaPID. This could be due to the segment being smaller than the 0.5cM threshold which we set on the IBD detection tool.

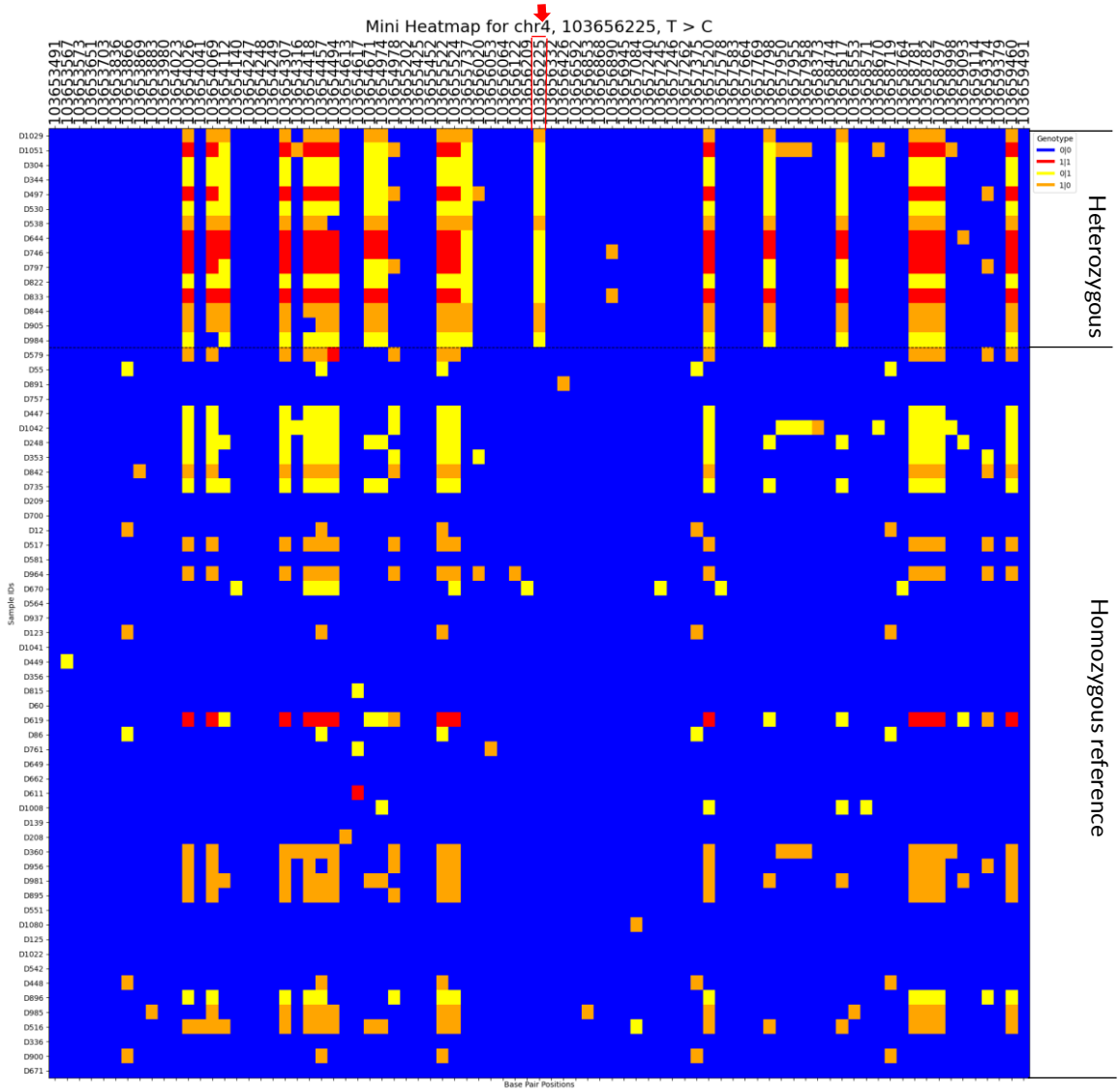

Figure 4.40: Heatmap showing the common IBD segment positions for the chromosome 4 variant TACR3 p.K286R*. This includes 15 heterozygous individuals and a random 50 homozygous reference individuals. The variant is indicated by the red arrow.*

Figure 4.41 represents the variant positions of the framework for the allele that carries the *TACR3* p.K286R variant (variant allele). This framework is of the individuals that carry the variant and ranges from the basepair position 103,635,509 to 104,041,448, approximately 405,000 basepairs. This equates to around 0.4cM, which is less than RaPID's 0.5cM threshold and hence why the tool was unable to detect it. It is possible that the heterozygotes of the dataset share a common ancestor since they have the same framework.

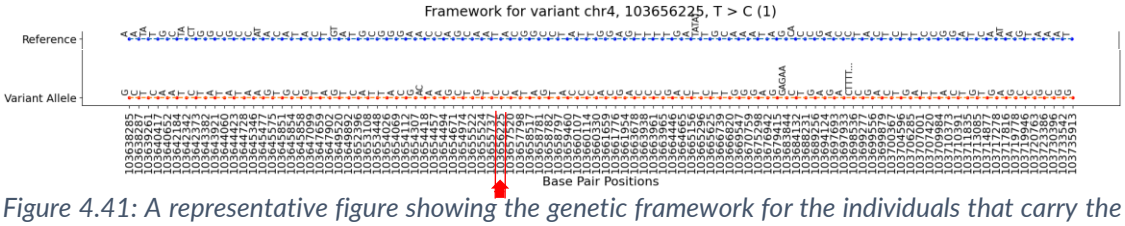

Figure 4.41: A representative figure showing the genetic framework for the individuals that carry the chromosome 4 variant TACR3 p.K286R. The variant is indicated by a red arrow. Abbreviated indels: deletion of ATATATGA at 103,665,296, insertion of TTTTCTTTT at 103,698,536.

4.3.6 Chromosome 11 Variants *HBB* p.T88P and *HBG2* p.H118R

The two chromosome 11 variants *HBB* p.T88P and *HBG2* p.H118R at positions 5,226,630 and 5,253,368 lie within 25,000 basepair positions of each other, and in most cases are inherited together (Kutlar et al., 1991), thus having the possibility of sharing an IBD segment. The first variant *HBB* p.T88P is a missense variant of a T nucleotide to a G. The *HBB* gene is involved in the production of the beta-globin protein, which is a subunit of haemoglobin (Hb). Variants in the *HBB* gene can either cause a decrease in production or produce a structural change of the beta-globin protein, both of which impairing the binding of oxygen (Greene et al., 2015). The Maltese population specifically contains multiple Hb variants, which include Hb F-Malta-I, Hb St. Luke's, Hb Long Island-Marseille and Hb Camperdown. The list also includes Hb Valletta, which is caused by the variant *HBB* p.T88P, leading to a substitution of threonine to proline in the beta chain. This was found in 34 Maltese and two Italian newborn babies. Although the variant causes a structural change in the protein, the individuals did not show any abnormal haematological results. In all of the newborn individuals, the Hb F-Malta-I variant was also detected, indicating close linkage between the two variants (Kutlar et al., 1991).

The aforementioned Hb F-Malta-I is caused by the variant *HBG2* p.H118R with a SNP of a T to C. The *HBG2* gene is involved in the production of gamma globin found in foetal Hb, which is then replaced with adult Hb at birth. Variants in this gene can either cause foetal Hb to remain throughout adulthood or produce new forms of foetal Hb (Greene et al., 2015), such as the one mentioned here. Hb F-Malta-I was first discovered in 12 of 658 infant cord blood samples, where on gel electrophoresis it moved slower than the normal foetal and adult

Hb (Cauchi et al., 1969). This was later investigated in 18 heterozygote newborns and 28 relatives older than two years of age. Fifteen of these relatives had Hb F-Malta-I, comprising 0.011% of all Hb in their blood. The normal foetal Hb levels in these 15 relatives, 11 other normal relatives and 50 normal adult controls were nearly the same (Altay et al., 1977).

Clinically both of the aforementioned variants do not cause any health issues. However, being almost exclusively found in the Maltese and tightly linked with each other, it is possible that they are part of an IBD segment and hence founder variants. Figure 4.42 shows the IBD segments that were detected for both of these variants, being close to each other. The dataset consisted of 20 and 19 heterozygous individuals for the variants *HBB* p.T88P and *HBG2* p.H118R respectively, of which in 16 an IBD segment was detected (80% and 84.21% of all individuals with the variant), with six homozygous reference individuals.

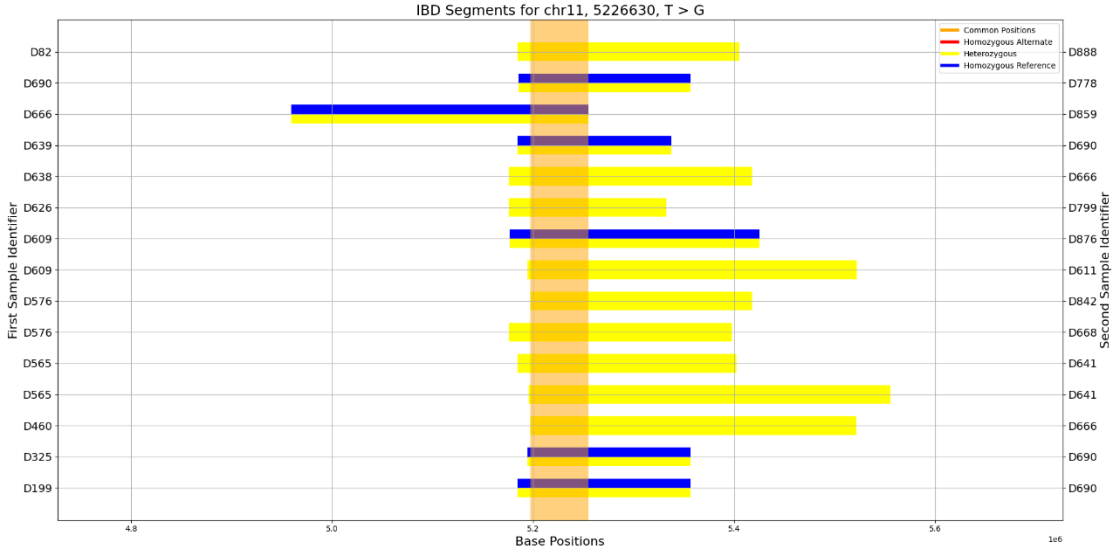

Figure 4.42: Horizontal bar plots showing the detected IBD segments for the chromosome 11 positions 5,226,630 and 5,253,368. This includes 16 heterozygous individuals for the variants HBB p.T88P and HBG2 p.H118R.

Figure 4.43 shows two heatmaps around the area of the variants which lies within the common position range of the IBD. This includes all heterozygous individuals and a random 50 homozygous reference individuals. The heatmaps show the presence of an IBD segment which is present on the allele opposite to the variants. This is found in homozygous reference and heterozygous individuals alike. However, it is possible that this is overlapping another undetected IBD segment which is present on the allele that carries both of these variants.

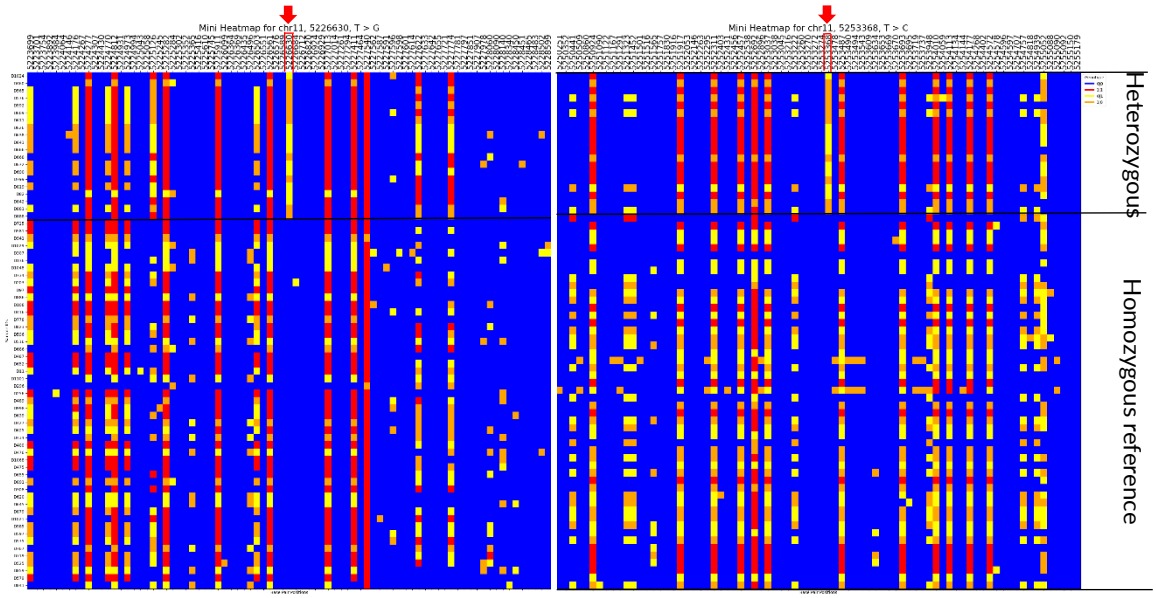

Figure 4.43: Heatmaps showing the common IBD segment positions for the chromosome 11 variants HBB p.T88P and HBG2 p.H118R. These include 20 and 19 heterozygotes for the variants respectively, together with a random 50 homozygous reference individuals.

The genetic framework of the variant allele generated from all individuals that carry both variants (Figure 4.44) confirms that the two variants are linked to each other. Being specific to the Maltese, these variants are most likely founder variants of the population. It is also possible that they arose within the Maltese population and therefore homozygous reference individuals with the ancestral framework on which the two variants arose are also being detected as having the same IBD segment.

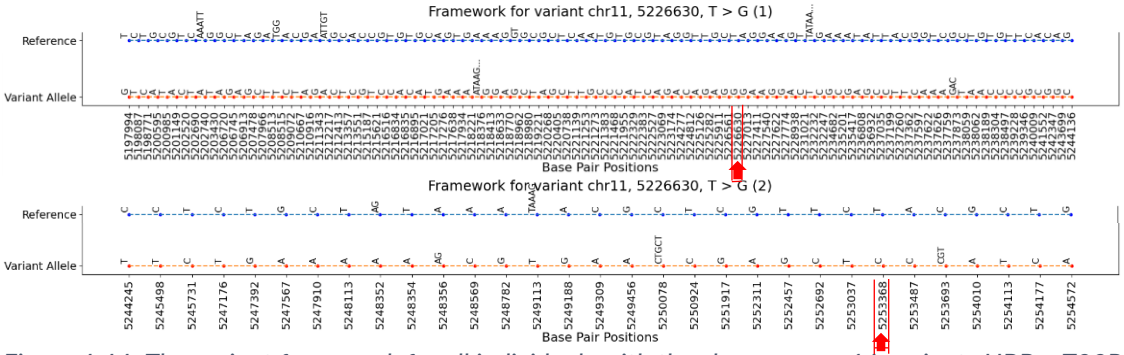

Figure 4.44: The variant framework for all individuals with the chromosome 11 variants HBB p.T88P and HBG2 p.H118R. The two variants form part of the same framework. The variants are indicated by red arrows. Abbreviated indel: insertion of TAAG at 5,218,376, deletion of ATAAAAA at 5,232,232.

4.3.7 Chromosome 19 Variant *NPHS1* p.R1160X

The chromosome 19 variant *NPHS1* p.R1160X is found in the nephrin producing gene. Nephrin is a transmembrane protein found in the glomerular

ultrafilter and plays an important role in signalling and structural functions. Defects in this protein can hinder the function of the glomerular filtration barrier, causing massive protein loss and congenital nephrotic syndrome (Khoshnoodi and Tryggvason, 2001). The disease onset occurs in newborns before three months of age and is characterized by proteinuria and oedema, as well as a large placenta in the mother. It ultimately progresses further to end-stage renal disease within two to three years (Ahvenainen et al., 1956). It has been primarily identified in the Finnish, who showcase many *NPHS1* variants (Huttunen, 1976; Norio, 1966). Later on, the pathogenic variant at position 35,831,056 which causes a stop codon in protein 1160 (p.R1160X), was identified by Lenkkeri et al. (1999) in five patients of Italian origin. The same variant was found in 13 Maltese individuals (from 11 families), suggesting a founder effect (Koziell et al., 2002).

The dataset consisted of 27 heterozygote individuals for this variant, two of which found to share an IBD segment running through the variant's position 5,226,630 (Figure 4.45). No IBD segment was detected in the remaining heterozygous samples.

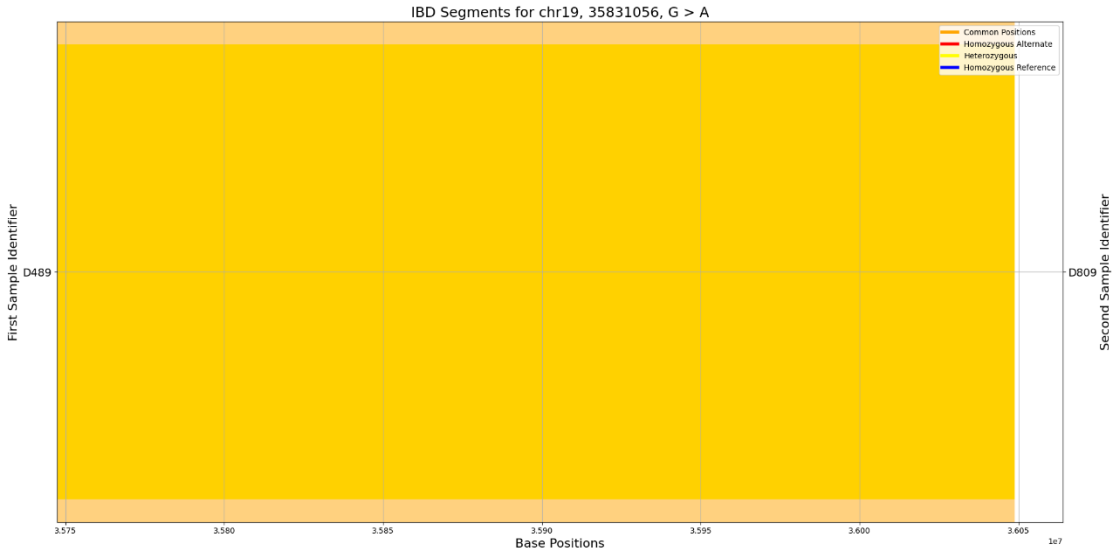

Figure 4.45: Horizontal bar plot showing the single IBD segment detected the chromosome 11 position 5,226,630. This includes 2 heterozygous individuals for the variant NPHS1 p.R1160X.

Figure 4.46 represents a heatmap around the area of the variant which lies within the common position range of the IBD segment. This includes all heterozygous individuals and a random 50 homozygous reference individuals.

However, this fails to represent an obvious pattern for the variant. It is possible that in such a region with a lot of genetic variation, as can be seen in Figure 4.46, recombination events may have contributed to the breaking down of any IBD segments. This would suggest that the variant is very old and therefore the IBD segment would be very small, unable to be detected by IBD detection tools. Although literature suggests the possibility that *NPHS1* p.R1160X is a founder variant, these results are inconclusive and this cannot be confirmed.

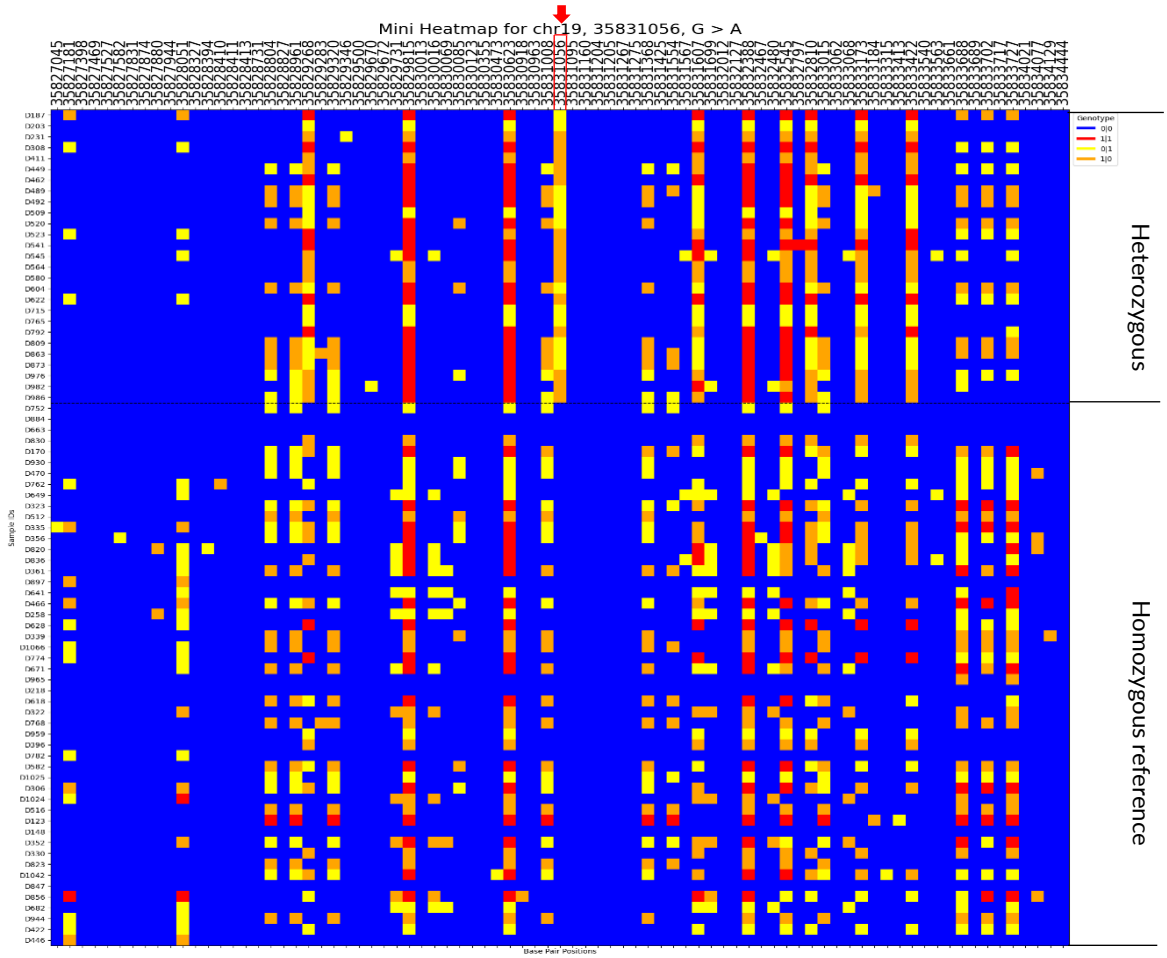

Figure 4.46: Heatmap showing the IBD segment position for the chromosome 19 variant NPHS1 p.R1160X. This includes 27 heterozygous individuals and a random 50 homozygous reference individuals. There is no obvious pattern for the variant, and thus its founder status cannot be concluded. The variant is indicated by the red arrow.

4.3.9 Undetected Variants

In five variants from the compiled list, no segments were detected by RaPID at both the 2cM and 0.5cM thresholds. Table 4.4 compiles these variants and includes information about the number of individuals and their zygosity in

our dataset. The inability to detect any IBD segments could be one of three reasons. First, they might not be founder variants of the Maltese population and thus are not part of a common IBD segment. Second, there are not enough individuals with the variants for RaPID to detect an IBD. In our dataset, four of the five listed variants have ten or less individuals and based on our results for other variants, not all of the individuals with the variant of study will exhibit the IBD. Third, it is possible that the variants remained undetected as they form part of a small IBD segment. With the lowest IBD segment threshold for RaPID being set to 0.5cM to avoid false positive results, smaller segments were filtered out. It is possible that some of these variants are very old and therefore their IBD got fragmented or shrunk with time due to genetic recombination events. Lower thresholds for RaPID should be able to capture such segments.

Table 4.4: A list of variants in which RaPID was unable to detect any IBD segments passing through their basepair position. Abbreviations: DHPR: Dihydropteridine reductase, CAAHD Congenital arthrogryposis with anterior horn cell disease, LCCS Lethal congenital contracture syndrome 1, HPFH: Hereditary persistence of foetal haemoglobin.

Variant Identifier	Gene	Disorder	Chr	Basepair position	DNA Nucleotide Change	No. of Individuals in Dataset	MAL_MAF
rs104893863	*QDPR*	DHPR deficiency	4	17,511,987	C>T	10 heterozygotes	0.007
rs1564162129	*GLE1*	CAAHD/LCCS	9	128,541,151	C>T	3 heterozygotes	0.003
rs753540084	*LRRK2*	Inborn genetic diseases	12	40,274,905	A>G	30 heterozygotes	0.019
rs770541847	*KISS1R*	Hypogonadotropic hypogonadism	19	919,929-919,965	GCGCGCCTACTGCAGTGAGGCCTTCCCCAGC>G	2 heterozygotes	0.001
rs267607202	*KLF1*	HPFH	19	12,885,368	T>A	3 heterozygotes	0.001

The variant in chromosome 4 *QDPR* p.G23D is responsible for the production of the dihydropteridine reductase enzyme. The enzyme is involved in the regeneration of BH_4, thus decrease in the enzyme causes decrease in BH_4 which leads to hyperphenylalaninaemia and phenylketonuria that causes brain and nerve damage. The variant is thought to originate from Mediterranean

populations (Dianzani et al., 1993), and has been found in the Maltese with a MAF of 0.007.

The *GLE1*, *LRRK2*, *KISS1R* and *KLF1* variants in Table 4.4 have so far mainly been reported in the Maltese, having a MAF of 0 in other populations. The *GLE1* p.S693F variant was found in a proband of Maltese ancestry in the homozygous state, whose non-consanguineous parents were considered healthy. Upon birth, the newborn male had abnormal facial characteristics and congenital contractures, typical of Lethal Arthrogryposis with Anterior Horn Cell Disease. The chromosome 9 variant at basepair position 128,541,151 was considered to be the best candidate for this, and the considered founder variant was found to be within a 7.7Mb run of homozygosity (Said et al., 2017). The novel *LRRK2* p.N618S variant with an amino acid change of asparagine to serine at position 618 was found in 73 cases of Parkinson's disease and 136 healthy controls from Malta, and is associated with an increased risk for the disease (Camilleri et al., 2015). The novel *KISS1R* p.Y190_A199del variant was identified in two Maltese individuals through a local study by Axiaq et. al., thought to be related to IHH (Reference: personal communication, presented at the European Society of Human Genetics 2024). The *KLF1* p.K288X variant was originally found in ten members of a Maltese family, and was linked to hereditary persistence of foetal haemoglobin in adults (Borg et al., 2010).

4.4 Outcomes of Founder Variant Analysis

Following the presentation of the results generated by the developed bioinformatics pipeline, six variants highly prevalent in the Maltese are suggested to be founder variants of the population. The heatmaps showed these variants as part of a genetic framework, which is shared among all of the individuals that have the variant, and thus suggesting a common ancestor. Four variants remained inconclusive upon interpretation of the results, while in another five variants, no IBD segments were detected at their location.

Identification of these founder variants allows for a more specific approach in genetic testing of the Maltese population. However, with the ever-changing populations and admixture of people from various cultures and ethnic groups, it is also important to put into perspective the population genomics of the last few years. As with the historical events that occurred hundreds or thousands of years ago, which shaped the genetic architecture of the current worldwide populations, ongoing real-time world events are continuously affecting genetic patterns. Discovering the genetic background of individuals through genealogical and ethnic history could enable opportunities to impact the communities from where they originate, making it possible to identify other community members who may be affected by genetic diseases. This could also have an effect on genetic testing, as unless a person is from the same ethnic group, targeted genetic testing can miss the real underlying disease-causing variant (Jain et al., 2021).

Following IBD detection with RaPID, visualisation of the IBD segments for the variants of interest was difficult at times. Having the presence of homozygous alternate individuals, such as in the *GNRHR* p.Q106R, *KISS1* p.X139fs and *KISS1* p.P81L variants, makes the identification of IBD segments easier, but these were missing for most of the variants. Homozygous alternate individuals simplify the genotype patterns and eliminate heterozygous noise. Visualisation of such segments is easier, and it is easier to infer that IBD segments between homozygous alternate individuals arise from a common ancestor. Such cases can also be used to identify large ROH. These are often indicative of a more recent common ancestor as such long segments represent IBD segments, hence the importance of detecting ROH. In fact, a ROH larger than 2cM was detected in chromosome 1 containing both *KISS1* variants and a ROH smaller than 0.5cM was detected in chromosome 4 with the *GNRHR* variant. Their presence made it easier to conclude the presence of an IBD segment.

Some difficulties were encountered as we attempted to automate the process of identifying founder variants using IBD segments. Ideally, a fully automated bioinformatics pipeline would immediately outline the founder status

of a variant, making for a more streamlined process. However, being a semi-automated pipeline, the developed method requires the manual interpretation of the generated outputs by the user. To identify the founder status of a variant, one has to manually look at the heatmaps and identify whether the variant forms part of an IBD segment, which may be undetectable by IBD detection tools because of their small size. This makes investigating a large quantity of variants inefficient. The horizontal bar plots outlining the common IBD segment positions and the genetic variant allele frameworks are otherwise automated and do not require manual interpretation.

While founder variants have been reported in several populations worldwide, there are only a limited number of ethnic groups and populations where founder variants have been thoroughly studied (Jain et al., 2021). Many publications and databases solely rely on the frequency estimations of the variants, without generating any proof or in-depth analysis. This could lead to many variants being mislabelled as founder variants for specific populations, which could have downstream effects on precision medicine, ancestry, bottleneck and migration studies. This could lead to misinterpretations of how populations would have evolved and interacted over time. It is important that such variants are thoroughly studied through various methods, such as the one presented here or through other methods, to confirm their founder status.

4.4 Summary

In this chapter, six of the most relevant IBD detection tools were tested across several power and accuracy metrics with the use of a benchmarking tool IBD Benchmark. This included testing of the tools across different genotype error rates of 0%, 0.01% and 0.1%, as well as multiple thresholds of IBD segment detection ranging from 50% to 100%. With the conclusion of RaPID being the best performing tool overall, optimisation of the tool was performed to obtain the best possible parameters. This was used to perform IBD detection for a list of variants of interest.

IBD detection at the 2cM threshold did not yield and IBD segments for the variants. However, this yielded a ROH for the two chromosome 1 variants *KISS1* p.X139fs and *KISS1* p.P81R. IBD detection was repeated at the 0.5cM threshold. From this, six variants from the compiled list were found to reside in five different IBD segments. These include the chromosome 1 variants *KISS1* p.X139fs and *KISS1* p.P81R which are part of the same IBD, the chromosome 4 variant *GNRHR* p.Q106R, the chromosome 4 variant *TACR3* p.K286R, and the chromosome 11 variants *HBB* p.T88P and *HBG2* p.H118R also part of the same IBD segment.

This is suggestive that the mentioned variants are founder variants of the Maltese population. A ROH was also detected in the chromosome 4 variant *GNRHR* p.Q106R. The chromosome 1 variants *CDCP2* p.P408RfsX46 and *KISS1* p.Q36R, the chromosome 2 variant *SPR* c.596-2A>G and the chromosome 19 variant *NPHS1* p.R1160X are inconclusive as these variants did not form part of the IBD segment that was detected by RaPID. No IBD segments were detected for the five other variants; the chromosome 4 variant *QDPR* p.G23D, the chromosome 9 variant *GLE1* p.S693F, the chromosome 12 variant *LRRK2* p.N618S and the chromosome 19 variants *KLF1* p.K288X and *KISS1R* p.Y190_A199del.

5. Conclusion

The primary aim of this study was to determine whether IBD segment analysis can be used to identify the founder status of a set of variants of interest. The first objective was to validate and compare the accuracy and performance of six IBD detection tools to choose the best performing one. After the optimisation of such tool and preprocessing of the Maltese dataset, the second objective was to construct a list of variants of interest and perform IBD detection in order to identify whether they can be classified as founder variants of the Maltese population. Many of the chosen variants are pathogenic, and have been reported to have a higher frequency in the Maltese when compared to other populations. The third objective was to develop a bioinformatics pipeline that aids the process of identifying founder variants and ROH. The method of selection of such tool, the steps involved in the identification of founder variants through IBD segment analysis and the developed bioinformatics pipeline were discussed in detail in the previous chapters.

5.1 Revisiting the Aims and Objectives

The first objective of this project was to compare the performance of six of the most used IBD detection tools in the last five years, hap-IBD, RaPID, RaPID-Query, FastSMC, RefinedIBD and IBDSeq. This was done with the first open-source benchmarking method called IBD Benchmark, which calculates the accuracy, length accuracy, length discrepancy, recall and power of the tools (Tang et al., 2022). A readily available phased sequencing European dataset (closest to Maltese) of chromosome 20 of 4,000 individuals provided by Tang et al. was used. With RaPID being the best performing and robust tool, especially with the introduction of genotype error rate, it was chosen to perform IBD analysis on a Maltese dataset. It was further optimised by testing out different window sizes and number of successes for the tool, with the window size of 5 and number of successes of 10 yielding the best results.

The second objective was to compile a list of variants of interest to the Maltese population, which consisted of 15 variants, and perform IBD detection. This was achieved by selecting variants which have a higher frequency in the Maltese when compared to other populations, many of which also reported to be pathogenic. This was followed by the IBD detection of the Maltese dataset using RaPID. This consisted of a single VCF of 1,076 genomes extracted from the Maltese population that formed part of the MAMI project (Attard et al., 2014), and was filtered down to 844 individuals after the removal of the relatives. Removal of these cases would ensure detection of IBD segments in unrelated individuals only and exclude any false positives. The dataset was divided on a chromosomal basis with VCFtools and genotype phasing was performed using Beagle, since RaPID needs both of these requirements.

The third objective was to develop a bioinformatics pipeline that automatically aids in the identification of founder variants. This was achieved with the creation of a Python script that identifies all IBD segments that involve the variants of interest. The start and end positions of the segments are plotted in a horizontal bar plot where the common range between them is highlighted. This range of positions is used to plot a heatmap containing homozygous reference, heterozygous and homozygous alternate individuals, if any. This can highlight any common genetic frameworks found in heterozygous and homozygous alternate individuals for the variant, by forming a genotypic pattern which will be similar amongst the individuals with the variant but absent from the rest. The variants involved within the IBD have to be on the allele that carries the variant of interest for the latter to be suggestive of a founder variant. The pipeline also handles indel cases where more than one variant is sometimes presented at the same position, hence filtering out the incorrect reads. The IBD variant framework is then plotted using all the individuals that share the segment and hence carry the variant of interest, within the common range highlighted earlier.

IBD detection was first performed at a 2cM threshold, filtering any segments below that measurement. None of the 15 variants of interest were

found to have an IBD with this threshold, however this yielded a ROH for the two chromosome 1 variants *KISS1* p.X139fs and *KISS1* p.P81R. IBD detection was repeated at the 0.5cM threshold. From this, six variants from the compiled list were found to reside in four different IBD segments. These include the chromosome 1 variants *KISS1* p.X139fs and *KISS1* p.P81R which are part of the same IBD segment, the chromosome 4 variant *GNRHR* p.Q106R, the chromosome 4 variant *TACR3* p.K286R, and the chromosome 11 variants *HBB* p.T88P and *HBG2* p.H118R also part of the same IBD segment. Being part of an IBD segment and forming part of a framework of variants, it is suggestive that the mentioned variants are founder variants of the Maltese population. At the 0.5cM threshold, a ROH was also detected in the chromosome 4 variant *GNRHR* p.Q106R.

The chromosome 2 variant *SPR* c.596-2A>G, the chromosome 19 variant *NPHS1* p.R1160X, and the chromosome 1 variants *CDCP2* p.P408RfsX46 and *KISS1* p.Q36R are inconclusive for founder variant status. These variants did not form part of the IBD segment that was detected by RaPID as such IBD segments resided on the opposite allele that did not carry the variant of interest.

No IBD segments were detected for the five other variants in the compiled list, which include the chromosome 4 variant *QDPR* p.G23D, the chromosome 9 variant *GLE1* p.S693F, the chromosome 12 variant *LRRK2* p.N618S and the chromosome 19 variants *KLF1* p.K288X and *KISS1R* p.Y190_A199del.

5.2 Limitations

Even though this project was successful in achieving its aims and objectives, the following limitations were encountered, which may have impacted the results.

One limitation of this project was in the testing of the IBD detection tools to find the best performing one to use. Upon compilation, more than 20 tools

were identified and it was not possible to test all of them due to the limited amount of time available and other objectives which we were aiming to reach. Since there is not any literature that has performed comparative analysis between them, it was necessary to select a few of the most used IBD detection tools in the last five years to test with IBD Benchmark and find the best performing tool out of them. With the testing of six IBD detection tools, the rest had to be excluded from project. Thus, there may be other tools which may perform better than our selected tool, RaPID. These include TBPWT (Freyman et al., 2021), Ilash (Shemirani et al., 2021), IBIS (Seidman et al., (2020) and TRUFFLE (Dimitromanolakis et al., 2019), among many others.

Another limitation of the project was related to ethics clearance for use of the data; the MAMI study data does not have ethics clearance for studies that may be related to ethnicity. By knowing the length of the IBD segment, one can estimate the age of the IBD and respective founder variant, as well as the most recent common ancestor. The longer the IBD segment, the more recent the common ancestor is. Tools such as DMLE+ (Reeve and Rannala, 2002) are able to infer such calculation. This in turn can be used to estimate a population's demographic history over time, including population size, bottlenecks and subsequent founder effects (Sticca et al., 2021). However, associating IBD segments or founder variants with other populations and age estimation inference of IBD segments could not be performed.

Following the previously mentioned limitation, identification of founder variants through IBD detection alone may not be the best method. Due to the limited resources available (limited only to VCFs), other methods of founder variant analysis were not possible. The use of IBD segment age estimation tools such DMLE+ would apply a possibly more accurate and efficient method. Such tools focus more on the variant of interest, using coalescent theory to trace the genealogical history of a variant and estimate how long ago it arose. This does not require the identification of large IBD segments, but rather looks at the variant of interest and surrounding haplotypes. Such tools also account for the demographic history of the population by using mutation rate, population size,

and generation time parameters, increasing the accuracy of the results. IBD detection tools do not make use of such functions, making it more difficult and less accurate in the identification of founder variants. The bioinformatics pipeline which uses the IBD detection method also requires the user to manually interpret the generated results as this cannot be completely automated. This would be inefficient for investigating a large quantity of variants and also possibly subject to interpreter bias.

Moreover, due to the limited amount of time and the type of resources required, confirmatory analysis of founder variants was not possible through wet lab analysis. One such method involves STR analysis, which uses short tandem repeats of two or more nucleotides that form a repetitive unit as biological markers (Fan and Chu, 2007). STRs are one of the most used molecular markers in genetic testing and previous studies have confirmed the relation of certain STR markers with specific disease-causing alleles. By taking STR markers close to the variant of interest, one can confirm the presence of a common allele through PCR amplification and sizing of fragments (Mejri et al., 2012).

Another limitation for the detection of IBD segments was the lack of homozygous alternate individuals for the majority of the variants. Such individuals have identifiable stretches of ROH, which have not been subjected to recombination events that break up IBD segments. They simplify the identification of genotype patterns and eliminate heterozygous noise, making it easier to identify IBD segments.

The script *variant_search.py* parses the *variants.txt* file, which is populated by the variants of interest in tab-delimited format. This includes the chromosome number, basepair position, reference allele and alternate allele, and uses this information to search for the variant within the VCF. In cases of indels, the representation of variants in the VCF may not always match the specified position and the respective reference and alternate alleles. On such example is the chromosome 1 *KISS1* variant at position 204,190,483 which is caused by a nucleotide change of CT>C. This marks a deletion of a T, which occurs 1 basepair position later (204,190,484). This is because indels are mapped to the last base

that was called during sequencing. These discrepancies can complicate the analysis as the variant basepair position, reference and alternate alleles inputted into the pipeline need to match exactly the entries in the VCF. The presence of the variant within the dataset should be confirmed beforehand as it may be represented differently than expected.

5.3 Future Work

Key achievements of this project include the identification of the best performing of the most used IBD detection tools, confirmation of a number of founder variants in the Maltese population and the development of a bioinformatics pipeline which can be used to study any variant for founder status using IBD segments. By successfully achieving the aims and objectives of this project, a solid foundation has been built on which further research can be done.

In the future, it would be beneficial to perform a comparative analysis on all available tools to ultimately find the best performing existing tool. This could be done through a benchmarking program like IBD Benchmark. Another option would be to test the tools on real data, and perform comparative analysis with the 15 variants that were tested in this project.

Following the identification of six possible founder variants in the Maltese population through IBD segment analysis, the developed bioinformatics pipeline and the readily generated IBD segments can be used to investigate other variants of interest and identify whether they reside within an IBD segment and are founder variants of the Maltese population. Moreover, the bioinformatics pipeline can also be used on other population datasets, given that the relevant data and IBD segments are generated for it.

There are other strategies that can be employed to identify founder variants with the use of IBD segments. One alternative way to do this would be to look solely for ROH that go through the basepair position of the variant, if available. The length of the ROH generally indicates the length of the IBD

segment, and if one is detected, it can be specifically used to search for the IBD segment in all of the carriers of the variant. By default, IBD detection tools are able to detect ROH segments alongside IBD segments. Another strategy which can be tested is by performing IBD detection on family groups that have the variant. Only one individual from each family was kept in our study as true IBD segments should reside in unrelated individuals, however family groups can be used to narrow down the location of any possible IBD segments which can then be searched for all of the carriers of the variant.

Confirmatory research of the founder variants identified in this project can be performed through STR analysis. This would require the selection of chromosomal specific STR markers from a database such as STRBase (Ruitberg et al., 2001), and through the use of fluorescently labelled flanking primers, amplifying them using PCR. Once amplified, the STR fragments are separated by size using capillary electrophoresis. The size of such fragments can be measured, corresponding to the number of repeat units. If a STR marker is always found to be associated with the adjacent variant, this contributes towards confirming the founder status of the variant (Almeida and Korch, 2004).

5.4 Final Remarks

In conclusion, having successfully met the key objectives of this study, numerous avenues for further research into founder variants have emerged. The findings of this study pave the way for further investigations into the use of IBD segments for the detection of founder variants and population studies. The bioinformatics pipeline that was developed in this project facilitates this process and should help researchers to identify more founder variants. Future research should aim to explore other avenues for founder variant analysis and provide deeper insights into their genetic impact to help refine our understanding of their role in disease and population genetics.

Appendix A

Table A1: The command line arguments and parameters used for the six IBD detection tools during IBD Benchmark analysis.

Tool	Parameters	Command Line Argument
RaPID	w=30 r = 10 s = 2 d = 2	./RaPID_v.1.7 -i <input_compressed_vcf_file> -g <genetic_map_file> -w 30 -r 10 -s 2 -d 2 -o <output_folder_name>
RaPID-Query	w=30 r=10 c=2 d=2 lm=700 dh=1.0 lmh=100 dg=10	./RaPID-Query_v1.0 -w 30 -r 10 -c 2 -d 2 -lm 700 -dh 1.0 -lmh 100 -dg 10 -m <output_file_name> -p <input_vcf_file> -q <input_vcf_file> -g <genetic_map>
hap-IBD	min-output=2	java -jar hap-ibd.jar gt=<input_vcf_file> out=<output_file_name> map=<genetic_map_file> min-output=2
FastSMC	--min_m 2	bcftools convert <input_vcf_file> --hapsample <output_file_name> ./FastSMC_exe --inFileRoot <input_hapsample_name> --outFileRoot <output_file_name> --decodingQuantFile decodingQuantities.gz --mode array --min_m 2 –segmentLength --hashing –perPairPosteriorMeans --perPairMAP --noConditionalAgeEstimates
RefinedIBD	length=2	java -jar refined-ibd.7Jan20.102.jar gt=<input_vcf_file> out=<output_file_name> map=<genetic_map_file> length=2
IBDSeq		sed '/^##/! s/|/\//g' INPUT.vcf > OUTPUT.vcf java -jar ibdseq.r1206.jar gt=<input_vcf_file> out=<output_file_name>

Table A2: The command line arguments and parameters used for the preparation of the MAMI VCF dataset and IBD detection with RaPID's optimal parameters.

Step	Parameters	Command Line Arguments
Unphasing	N/A	sed '/^##/! s/|/\//g' <input_vcf_file> > <output_vcf_file>
Filtering of Relatives	--recode	vcftools --vcf <input_vcf_file> --<input_relatives_ids.txt> --recode --out <output_vcf_file>
Chromosomal Extraction	--chr <chr_number> --recode	vcftools --vcf <input_vcf_file> --chr <chr_number> --recode --out <output_vcf_file>
Phasing	nthreads=20 window=10	java -jar Beagle.01Mar24.d36.jar gt=<input_vcf_file> map=<plink_genetic_map> nthreads=20 window=10 out=<output_vcf_file>
RaPID IBD Detection	-w 5 -r 10 -s 2 -d 2	./Rapid_v.1.7 -i <input_vcf_file> -g <map_file> -o <output_ibd_file> -w 5 -r 10 -s 2 -d 2

Appendix B

This appendix includes all the extra generated plots and figures that did not make the final dissertation writeup.

IBD Benchmark IBD Detection Tool Plots

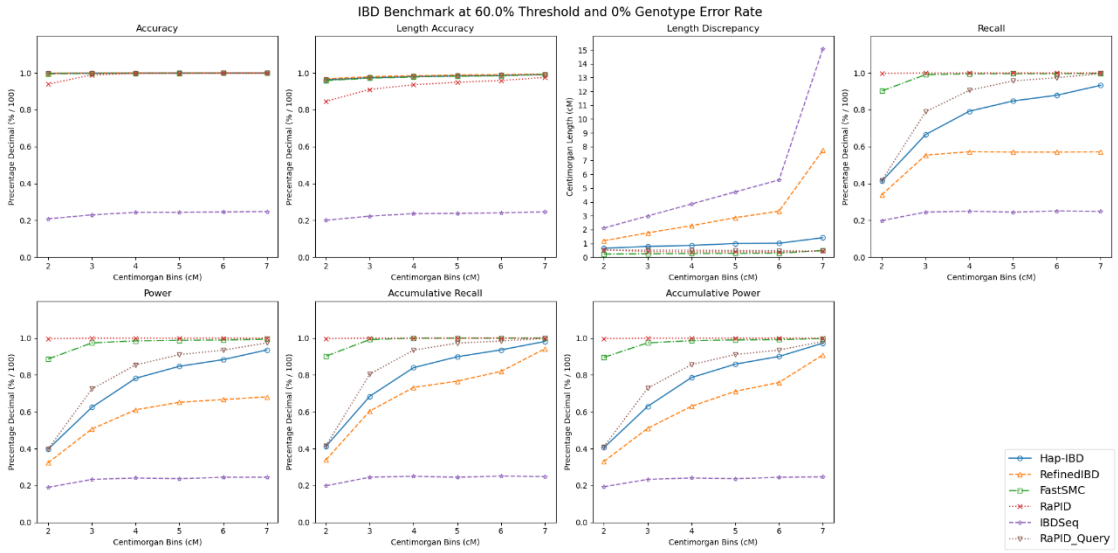

Figure B1: IBD Benchmark results of the IBD detection tools at 60% threshold and 0% genotype error rate.

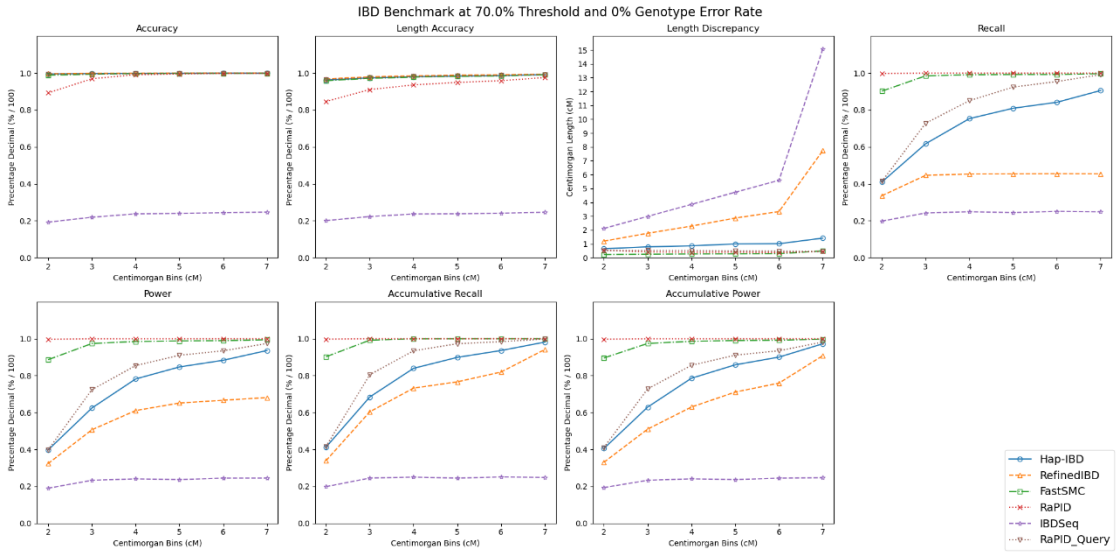

Figure B2: IBD Benchmark results of the IBD detection tools at 70% threshold and 0% genotype error rate.

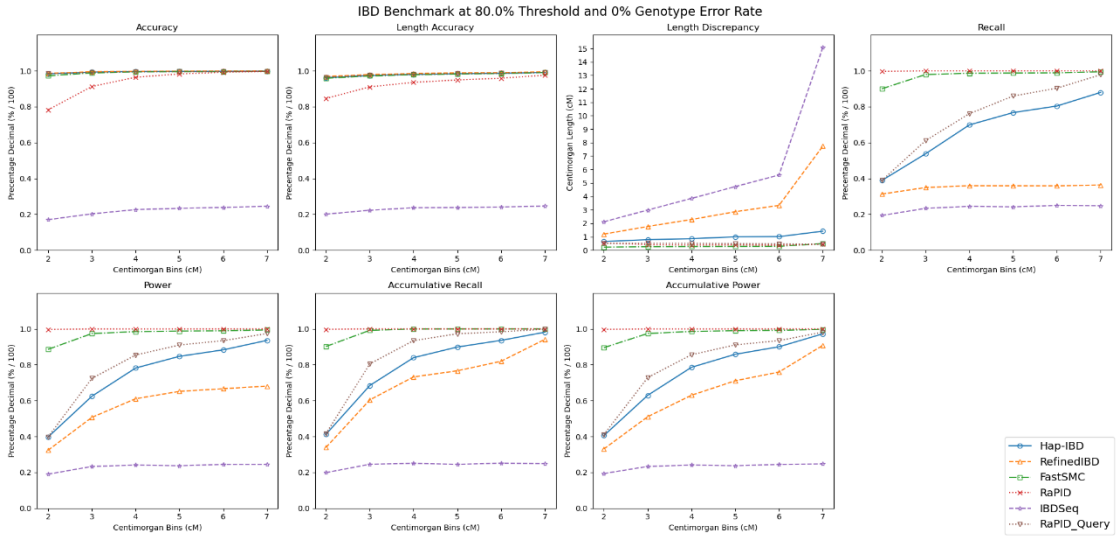

Figure B3: IBD Benchmark results of the IBD detection tools at 80% threshold and 0% genotype error rate.

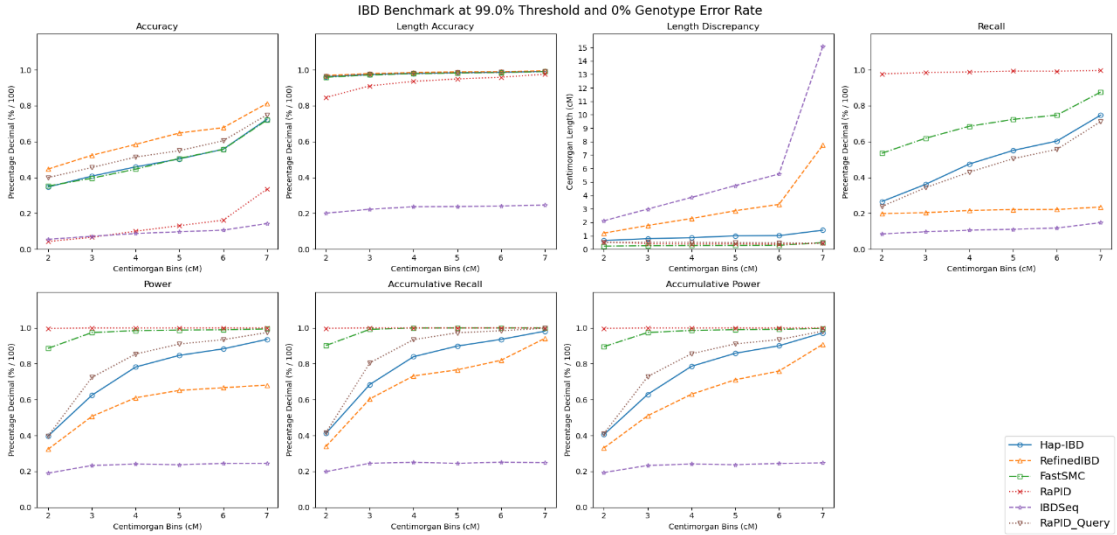

Figure B4: IBD Benchmark results of the IBD detection tools at 99% threshold and 0% genotype error rate.

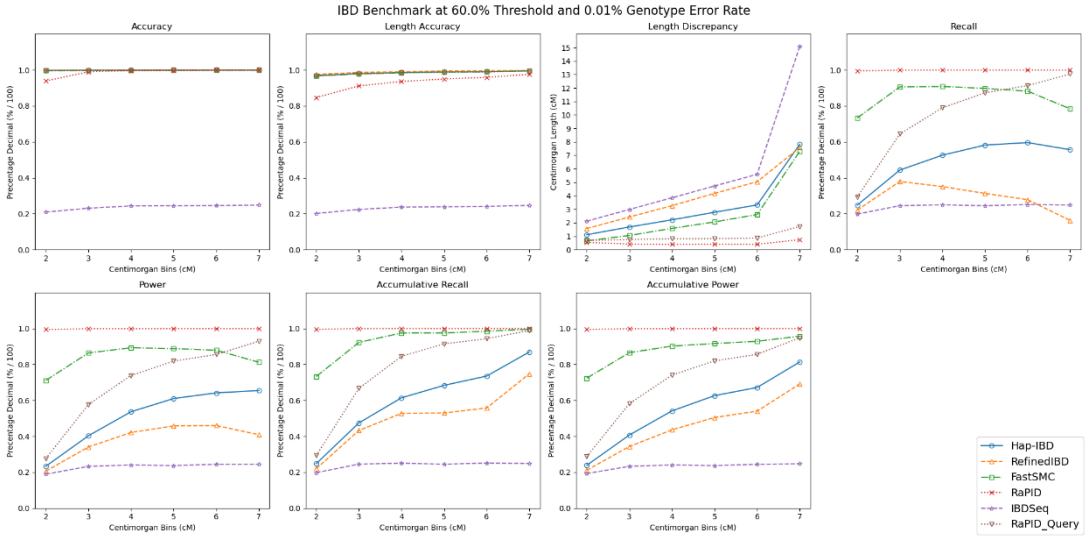

Figure B5: IBD Benchmark results of the IBD detection tools at 60% threshold and 0.01% genotype error rate.

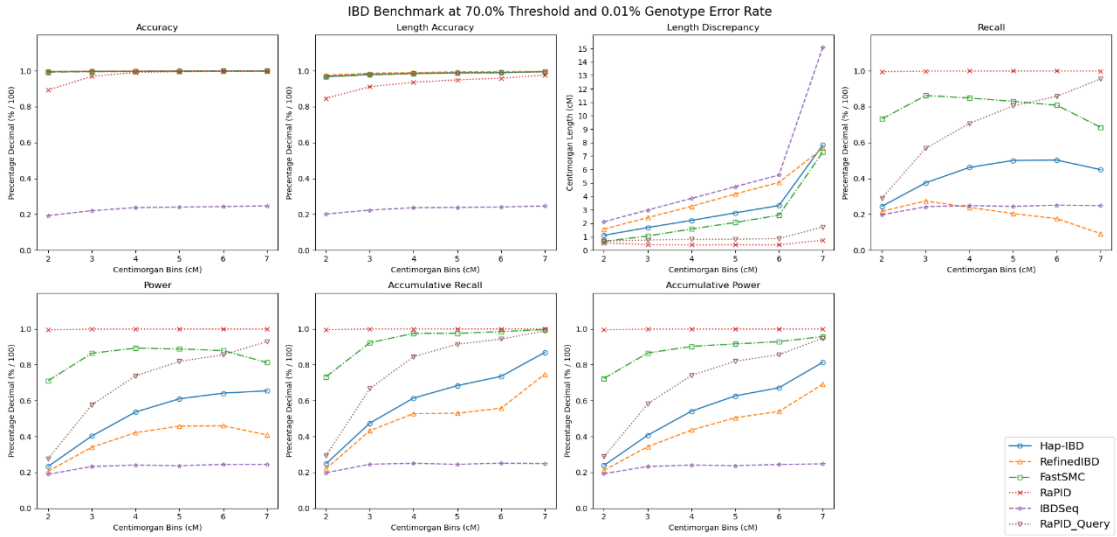

Figure B6: IBD Benchmark results of the IBD detection tools at 70% threshold and 0.01% genotype error rate.

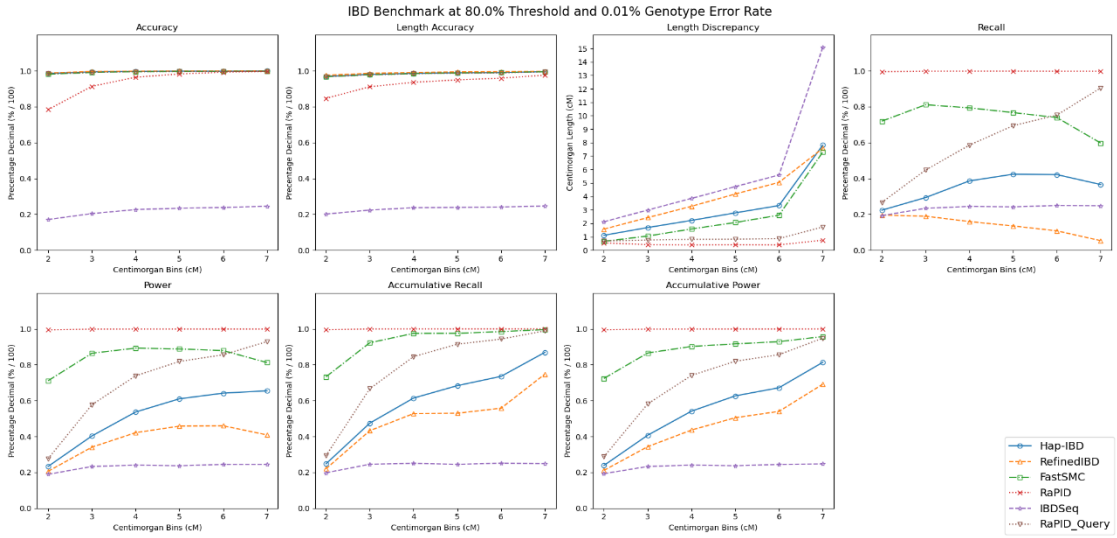

Figure B7: IBD Benchmark results of the IBD detection tools at 80% threshold and 0.01% genotype error rate.

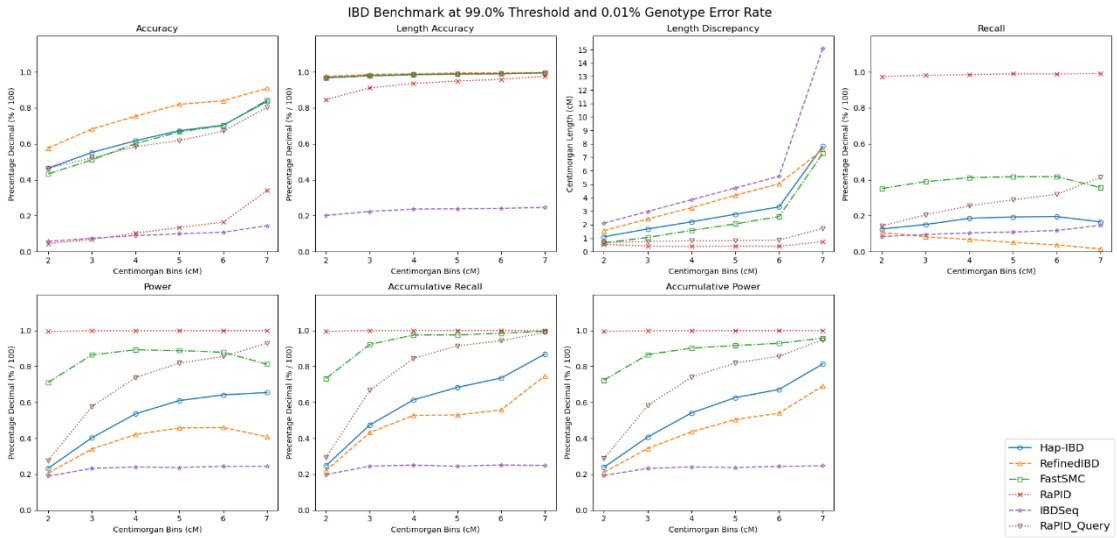

Figure B8: IBD Benchmark results of the IBD detection tools at 99% threshold and 0.01% genotype error rate.

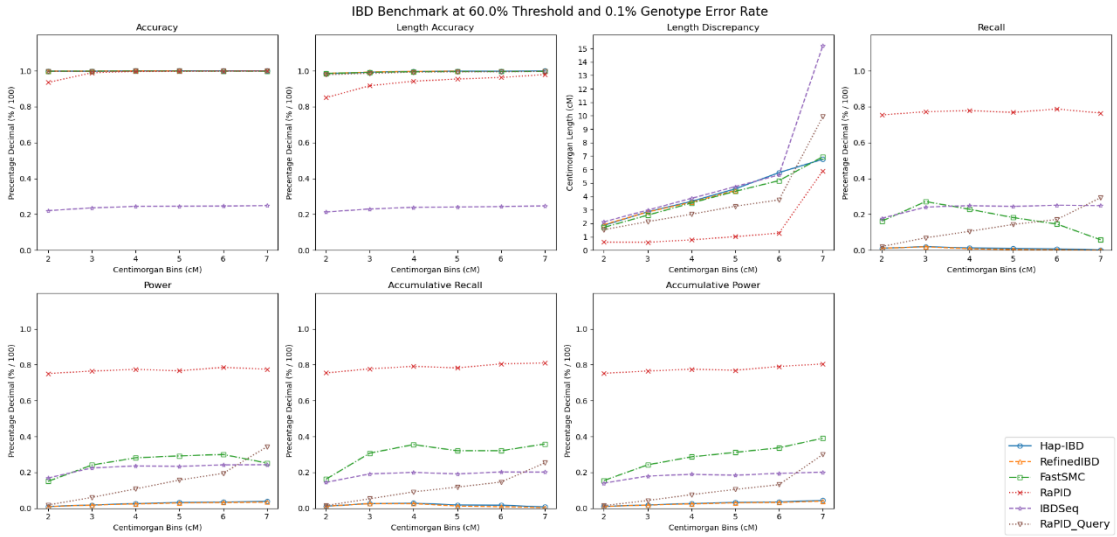

Figure B9: IBD Benchmark results of the IBD detection tools at 60% threshold and 0.1% genotype error rate.

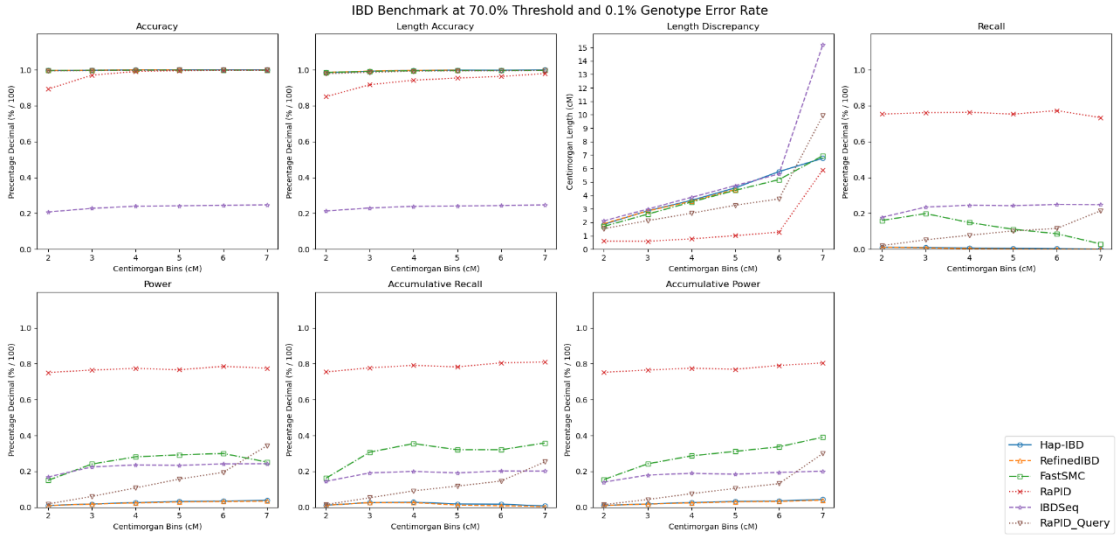

Figure B10: IBD Benchmark results of the IBD detection tools at 70% threshold and 0.1% genotype error rate.

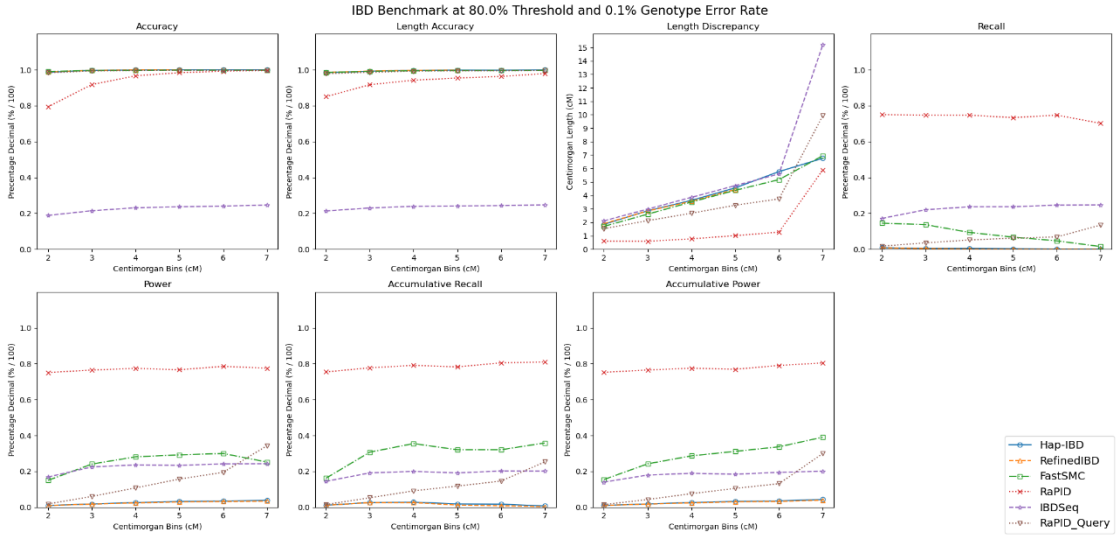

Figure B11: IBD Benchmark results of the IBD detection tools at 80% threshold and 0.1% genotype error rate.

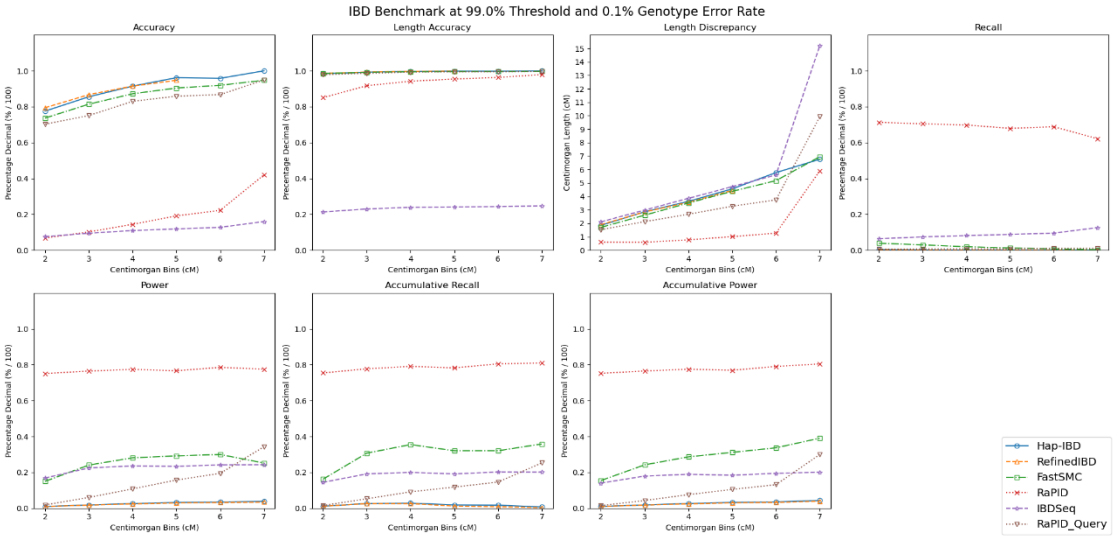

Figure B12: IBD Benchmark results of the IBD detection tools at 99% threshold and 0.1% genotype error rate.

IBD Benchmark RaPID Optimisation Plots

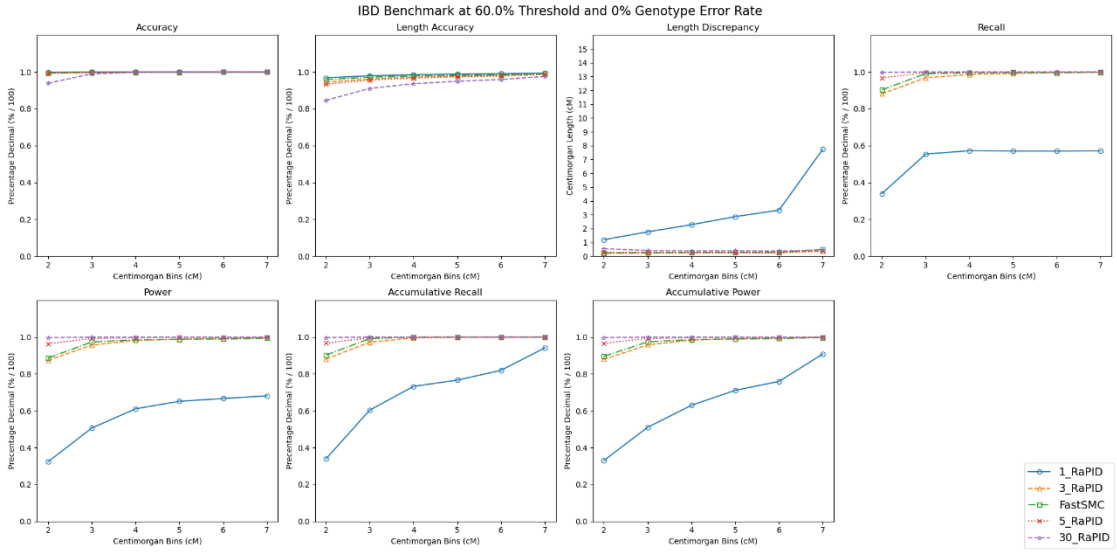

Figure B13: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 60% threshold and 0% genotype error rate.

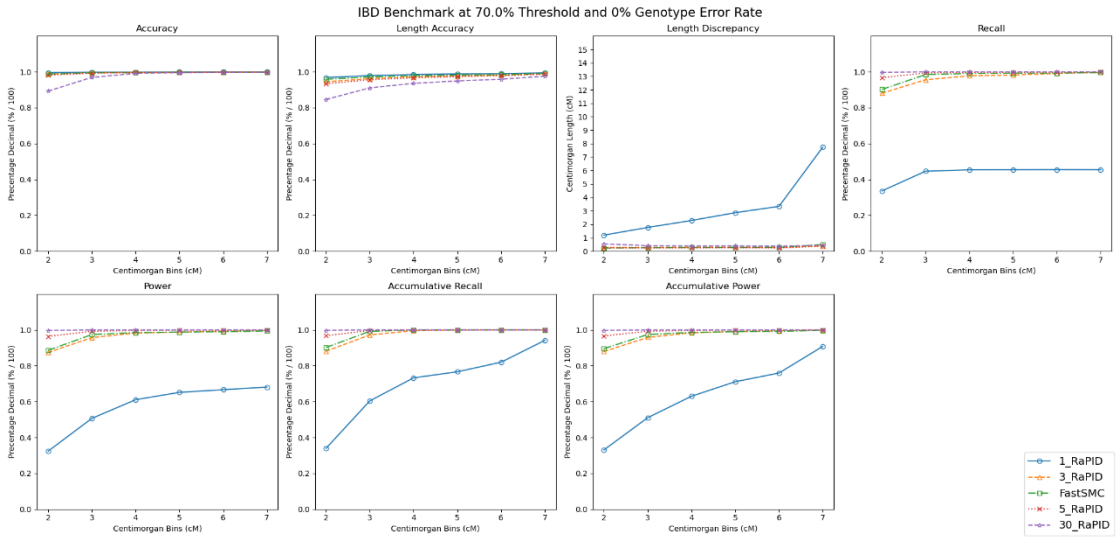

Figure B14: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 70% threshold and 0% genotype error rate.

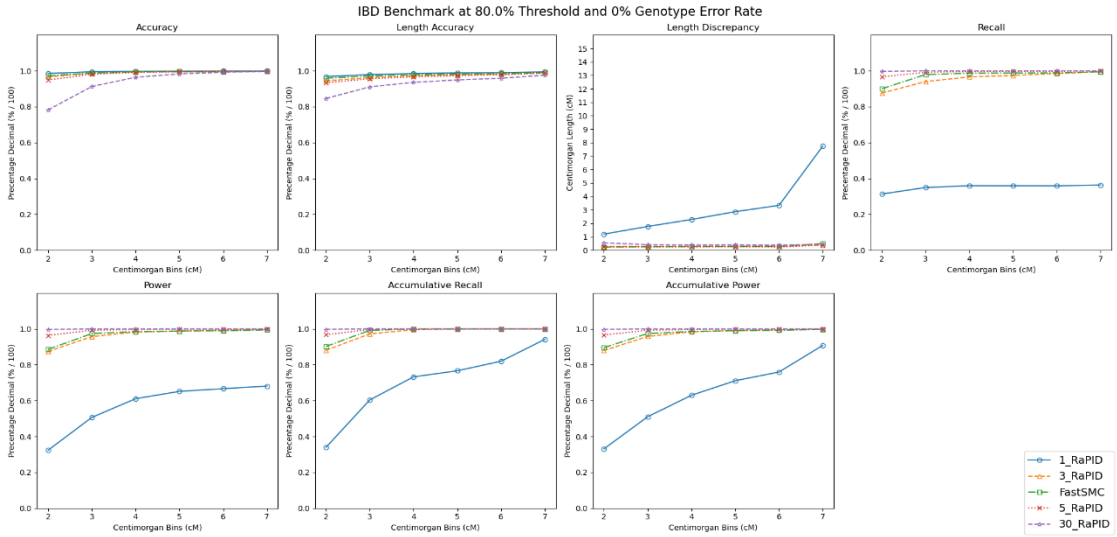

Figure B15: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 80% threshold and 0% genotype error rate.

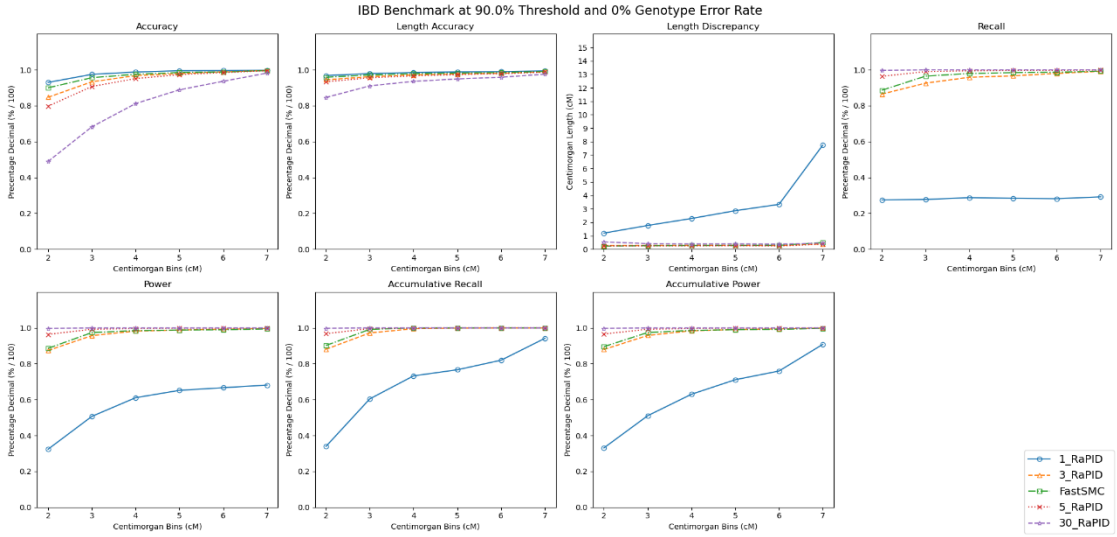

Figure B16: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 90% threshold and 0% genotype error rate.

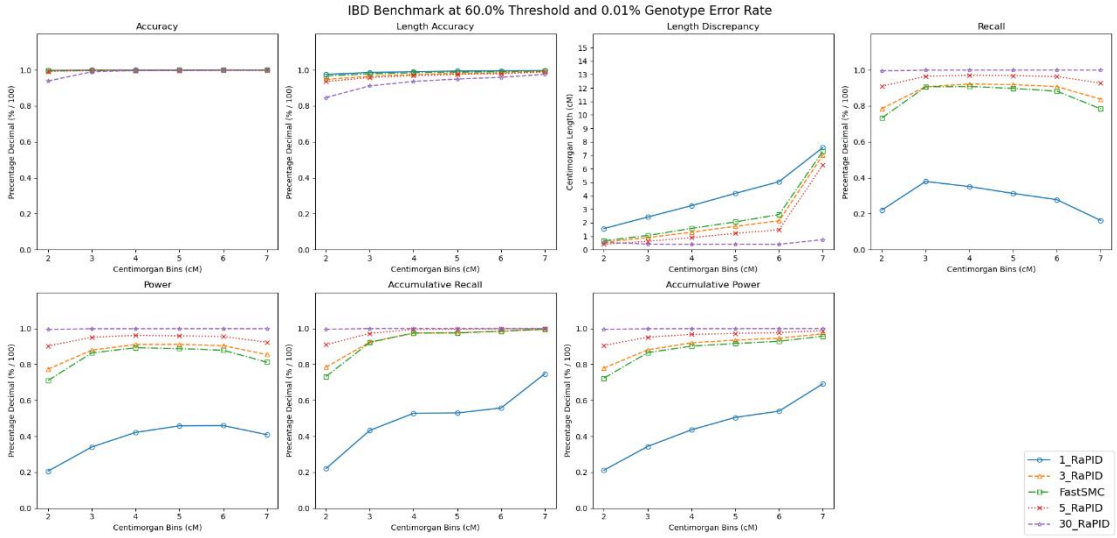

Figure B17: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 60% threshold and 0.01% genotype error rate.

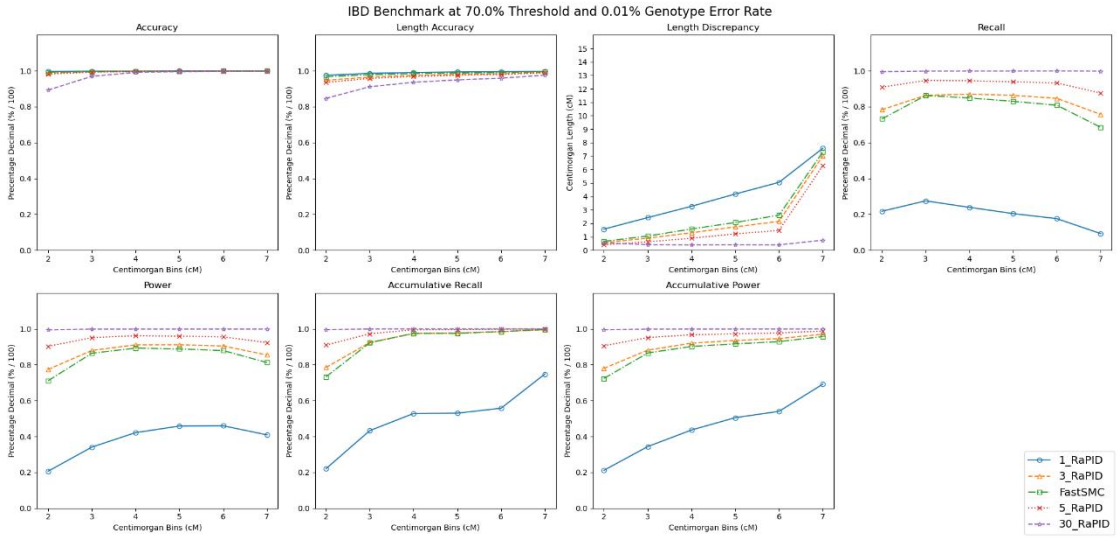

Figure B18: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 70% threshold and 0.01% genotype error rate.

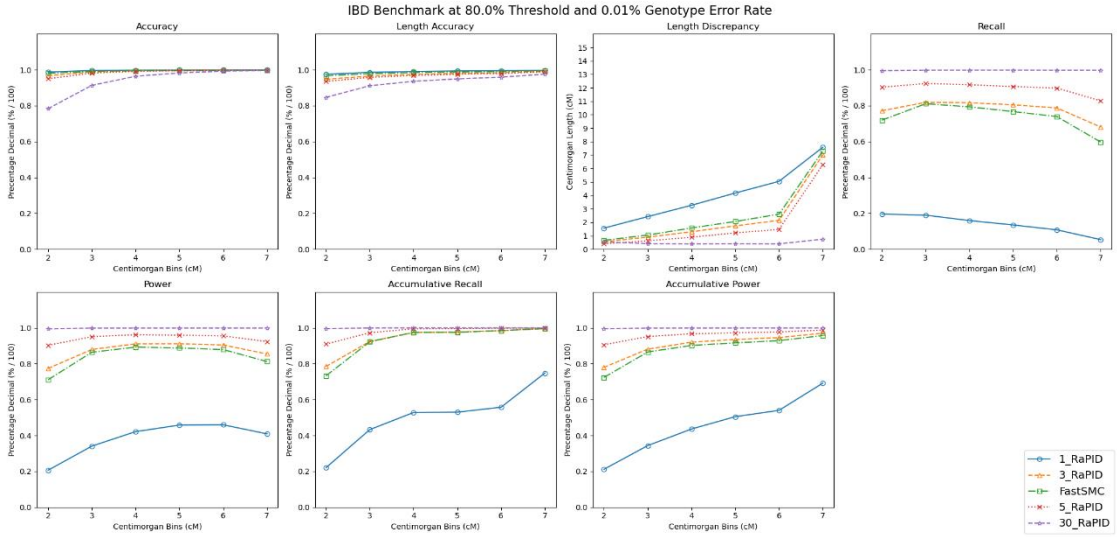

Figure B19: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 80% threshold and 0.01% genotype error rate.

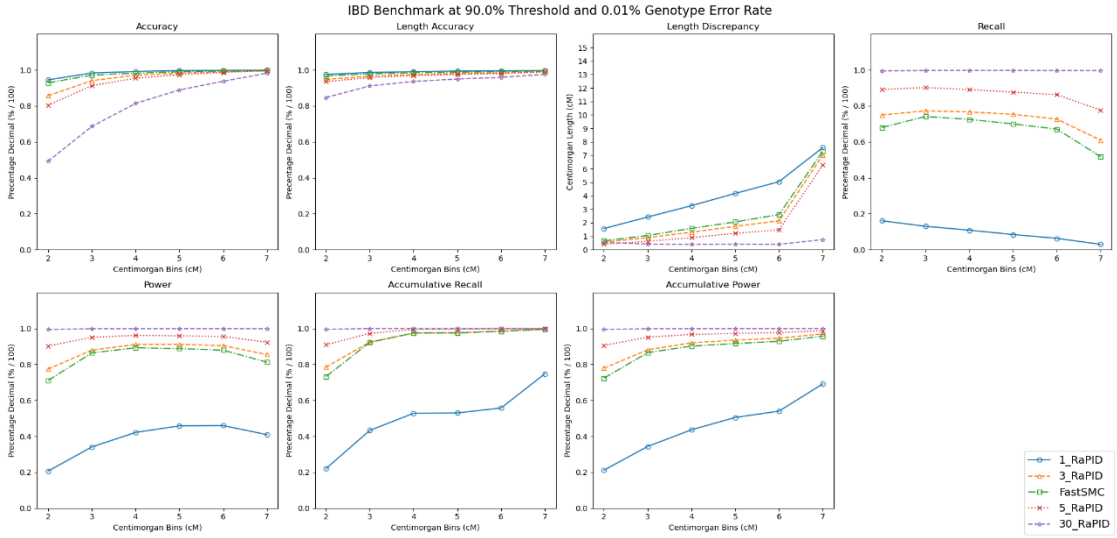

Figure B20: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 90% threshold and 0.01% genotype error rate.

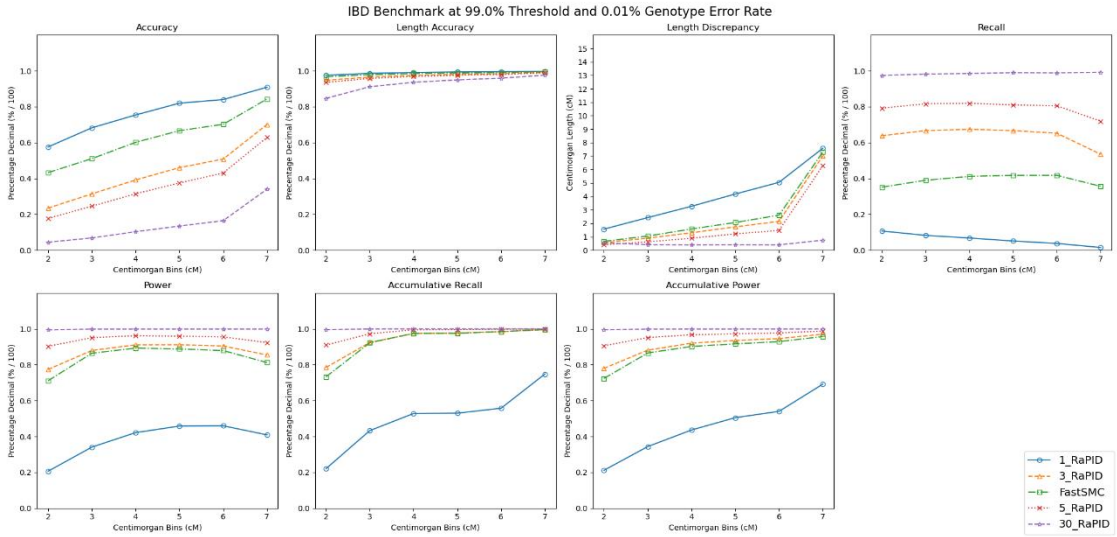

Figure B21: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 99% threshold and 0.01% genotype error rate.

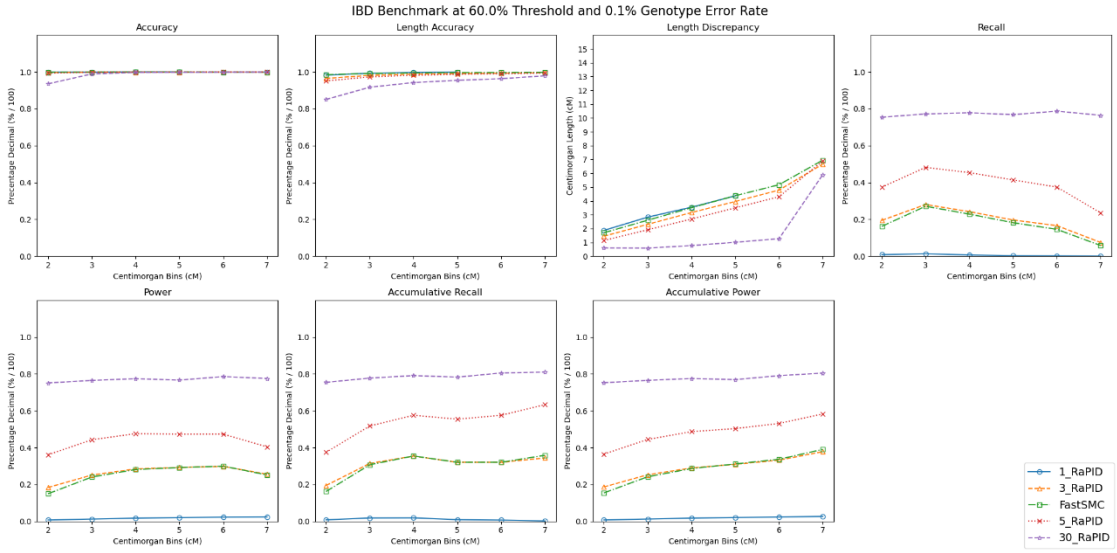

Figure B22: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 60% threshold and 0.1% genotype error rate.

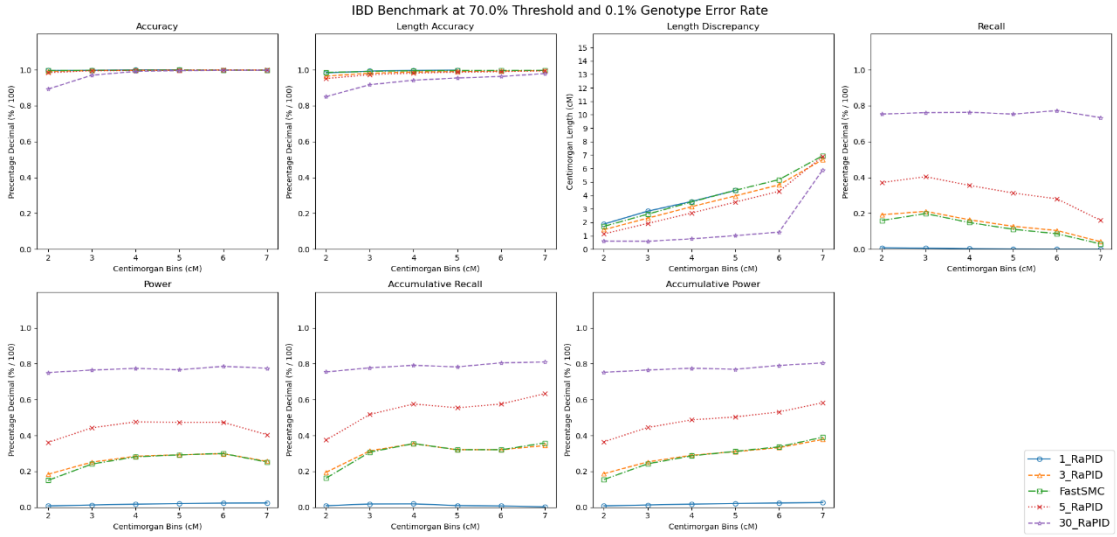

Figure B23: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 70% threshold and 0.1% genotype error rate.

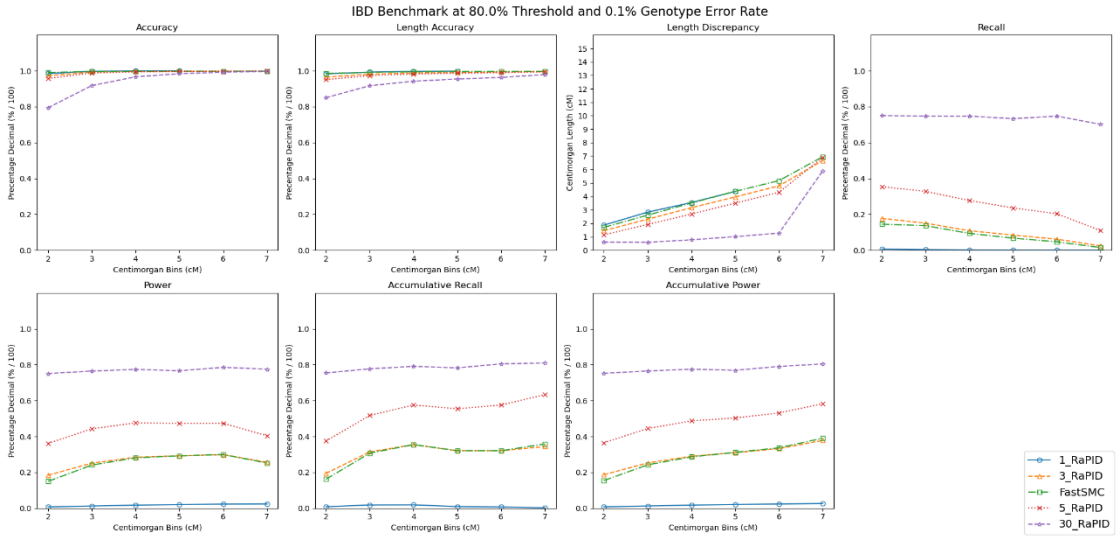

Figure B24: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 80% threshold and 0.1% genotype error rate.

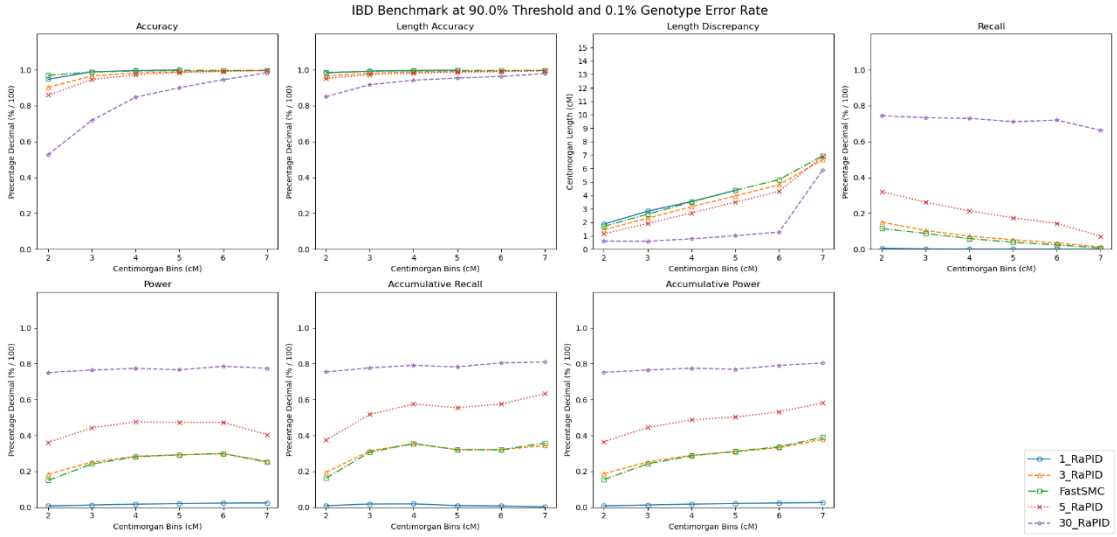

Figure B25: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 90% threshold and 0.1% genotype error rate.

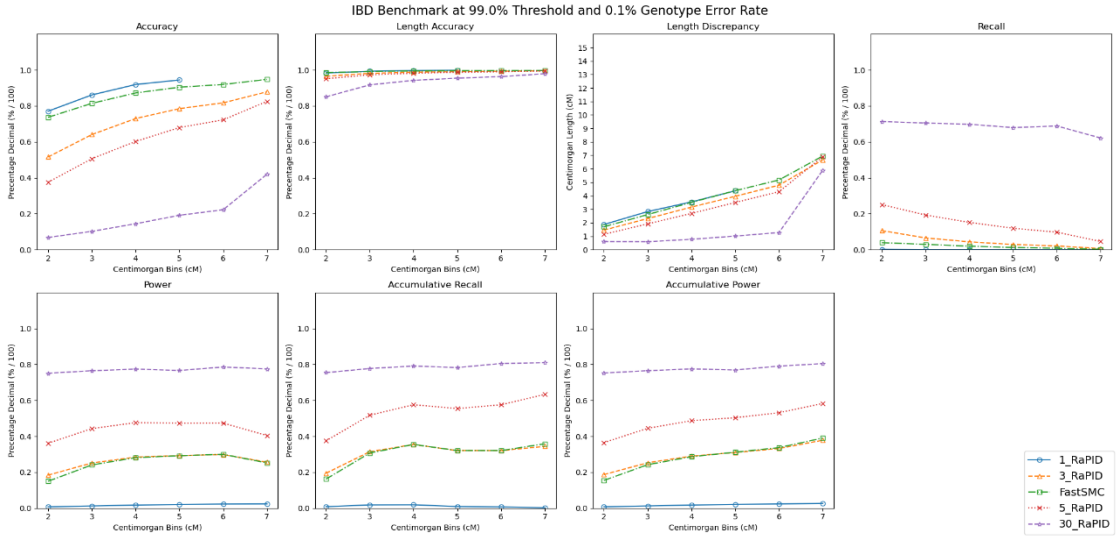

Figure B26: IBD Benchmark results of the RaPID's different window sizes of 1, 3, 5 and 30, and FastSMC at 100% threshold and 0.1% genotype error rate.

Variant Results

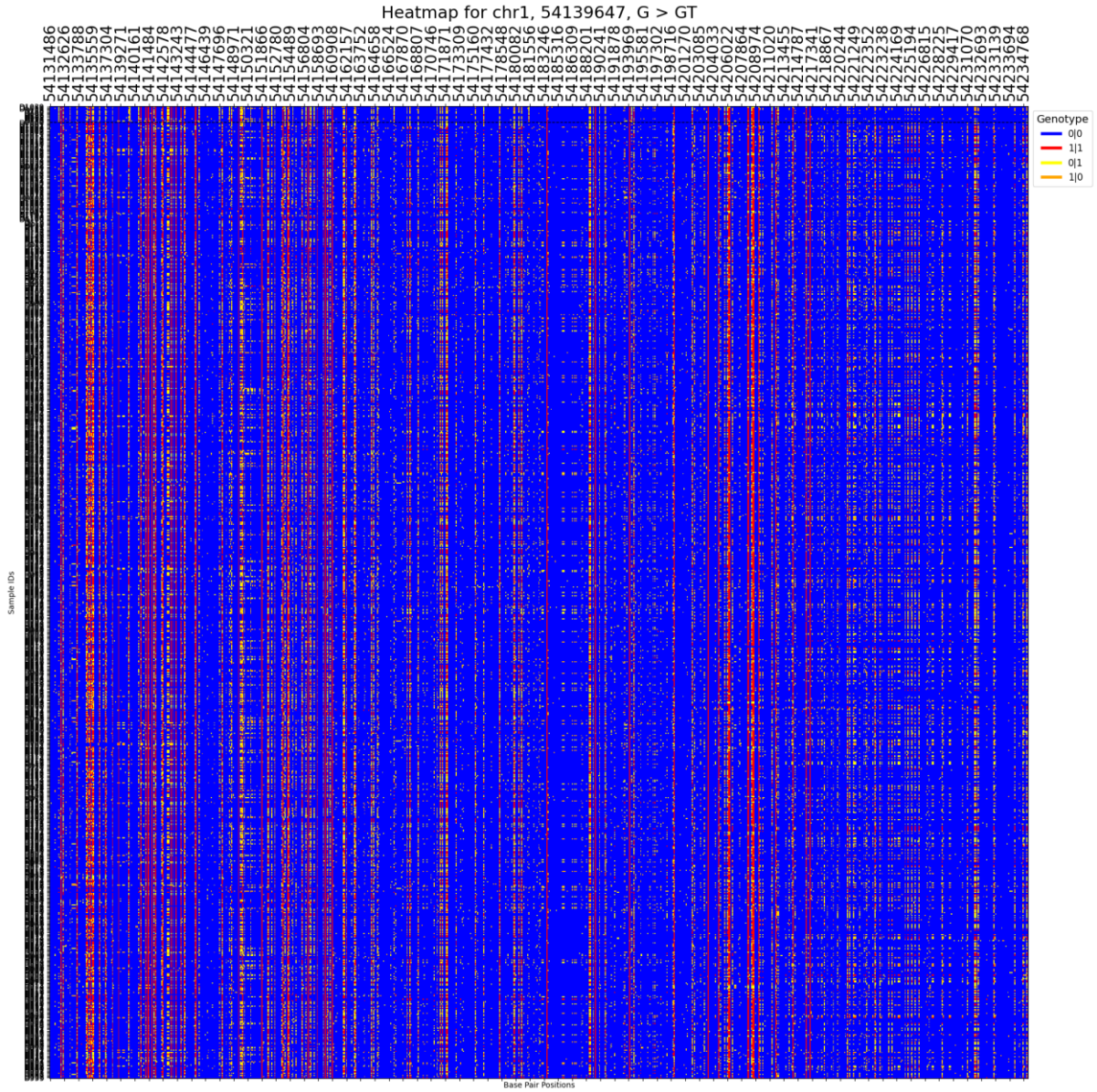

Figure B27: Full heatmap showing the common IBD segment positions for the chromosome 1 variant CDCP2 p.P408RfsX46.

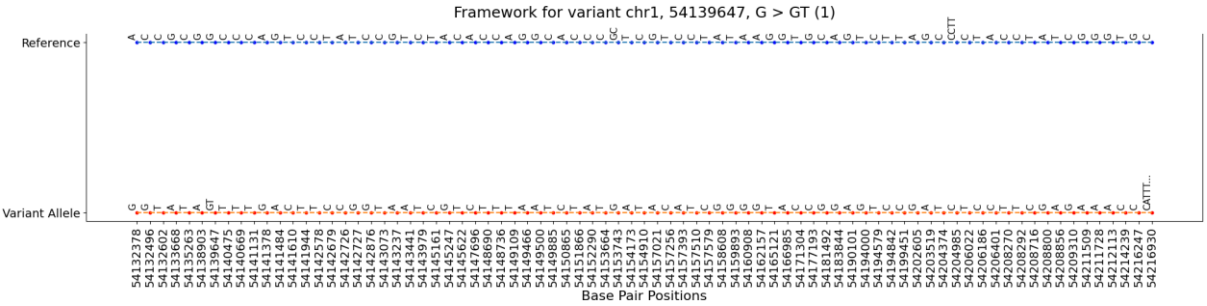

Figure B28: The genetic framework for the individuals that carry the chromosome 1 variant CDCP2 p.P408RfsX46.

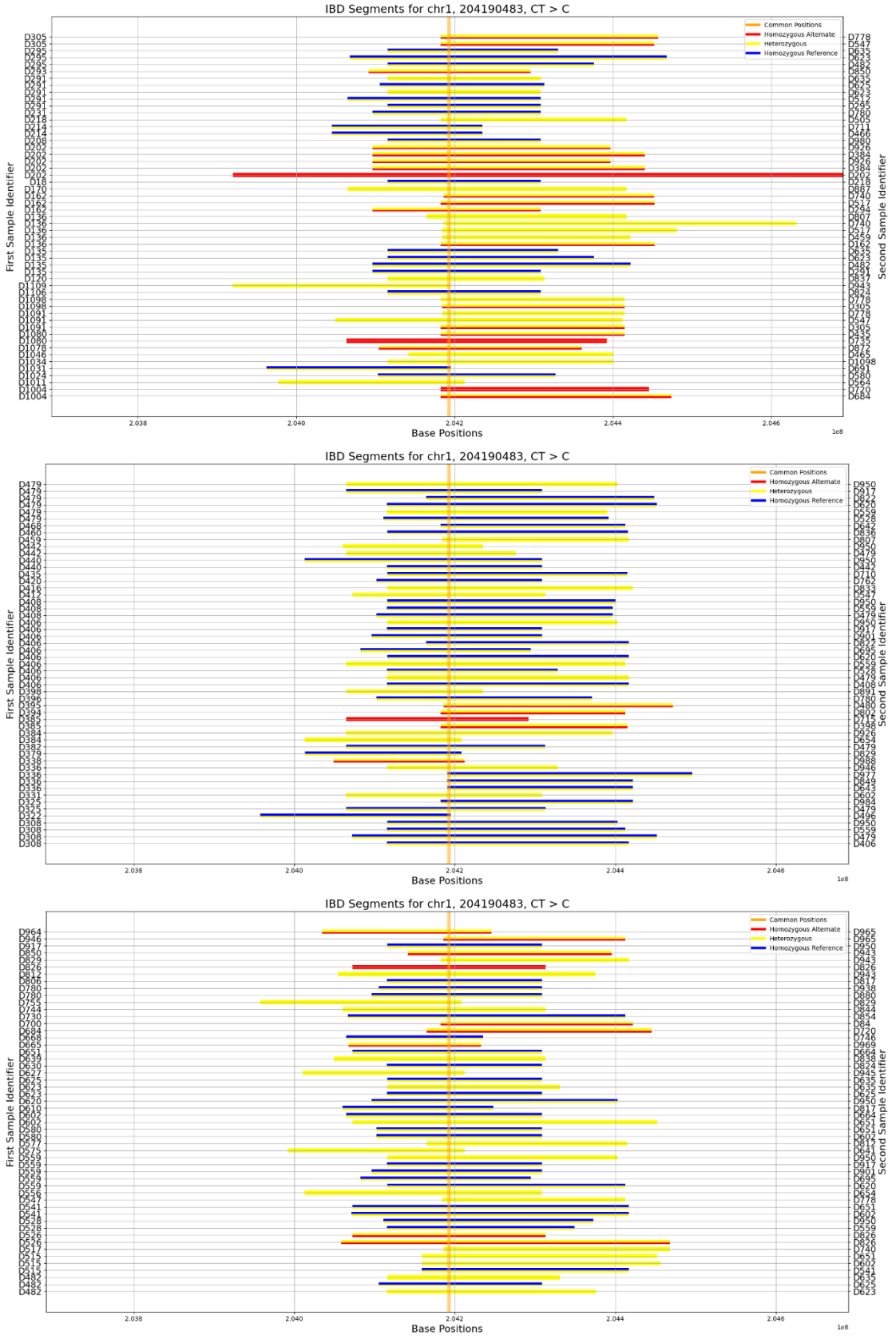

Figure B29: All horizonal bar plots showing the detected IBD segments for chromosome 1 position 204,190,483.

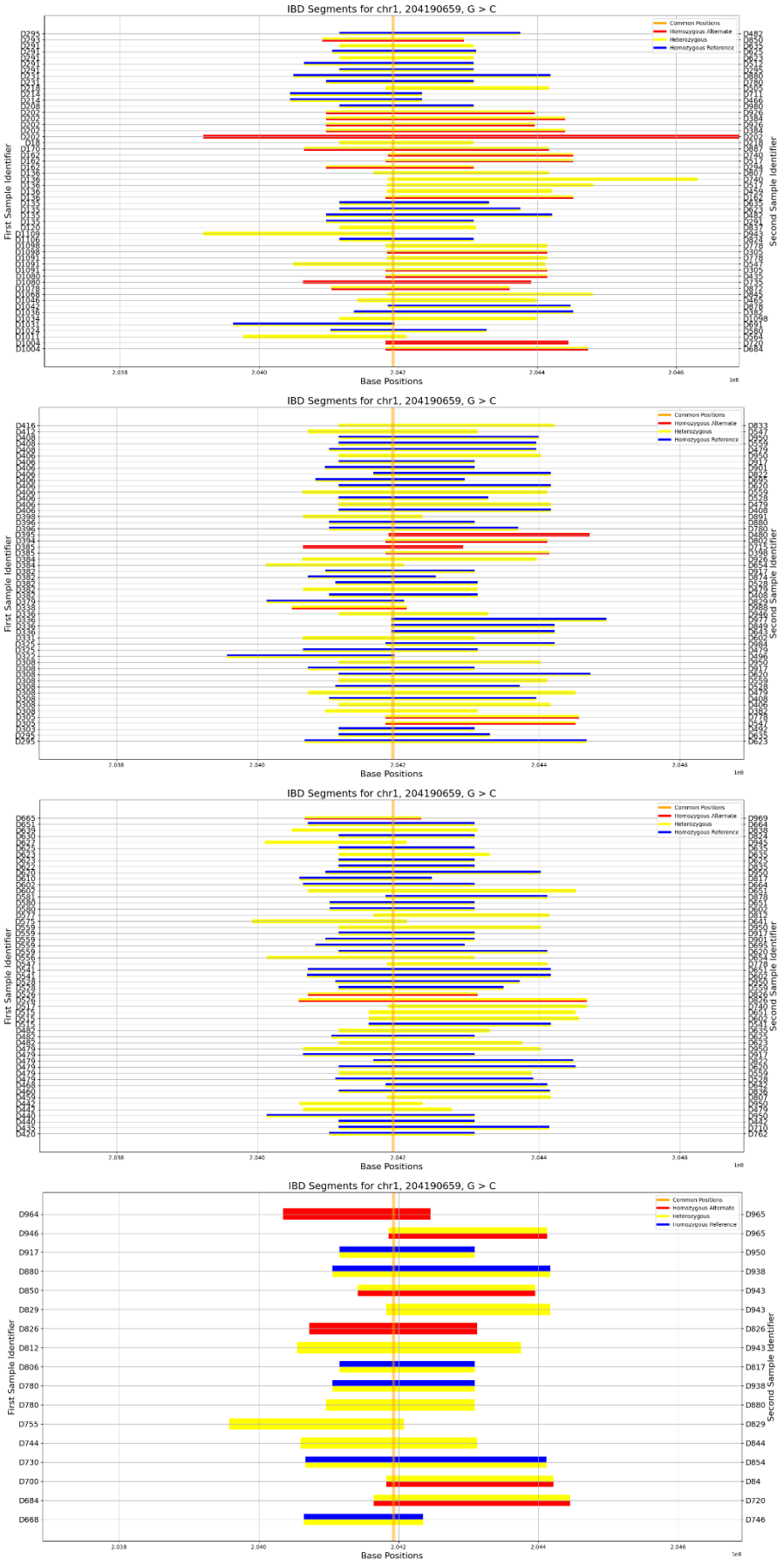

Figure B30: All horizonal bar plots showing the detected IBD segments for chromosome 1 position 204,190,659.

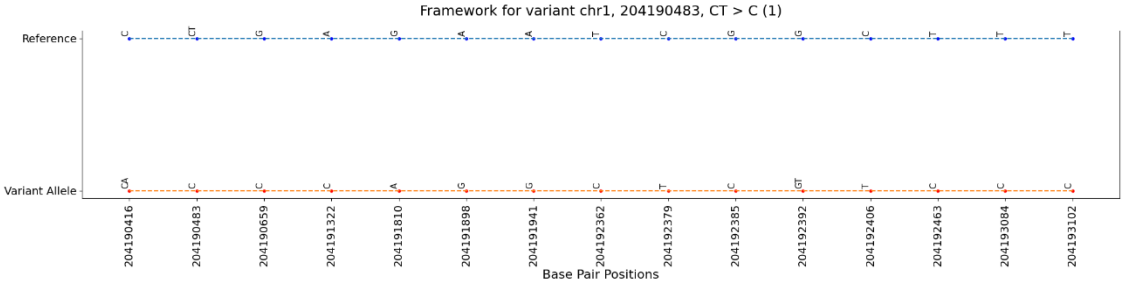

Figure B31: The genetic framework for the individuals that carry the chromosome 1 variant KISS1 p.X139fs. This is the same genetic framework for the chromosome 1 variant KISS1 p.P81R.

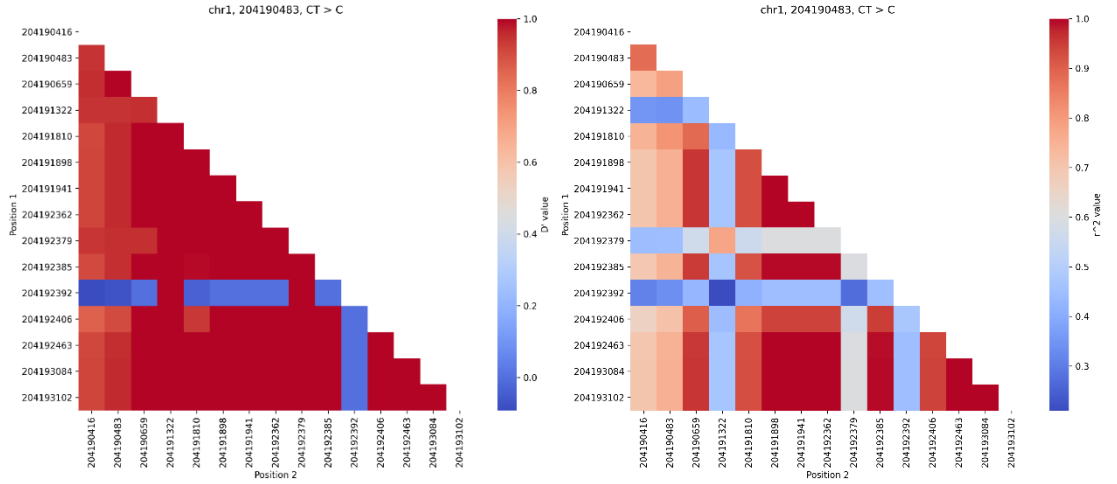

Figure B32: LD plots for the variant framework of the chromosome 1 variants KISS1 p.X139fs and KISS1 p.P81R, showing the high LD between the variants.

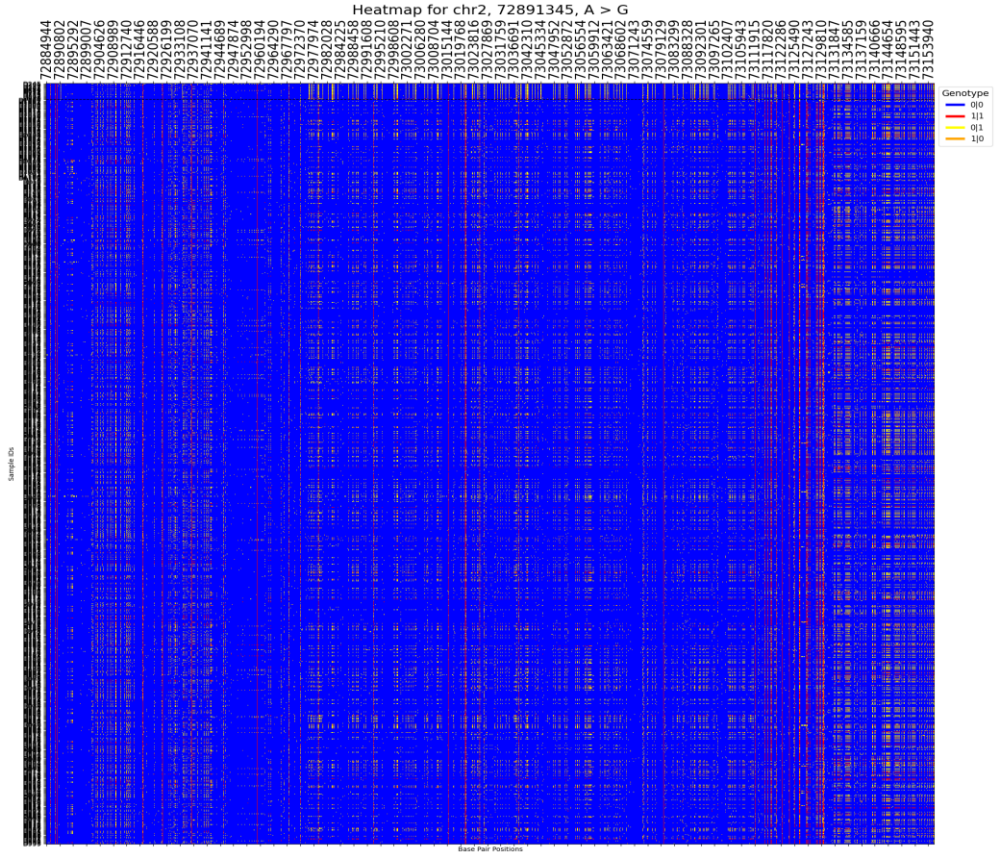

Figure B33: Full heatmap showing the common IBD segment positions for the chromosome 2 variant SPR c.596-2A>G.

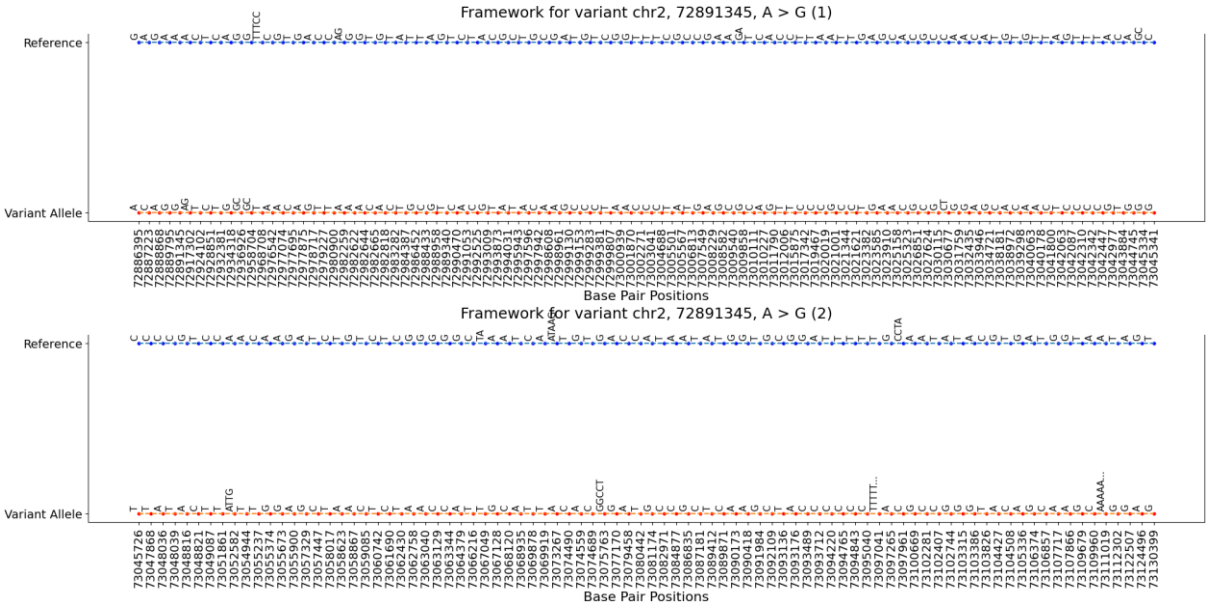

Figure B34: The genetic framework for the individuals that carry the chromosome 2 variant SPR c.596-2A>G.

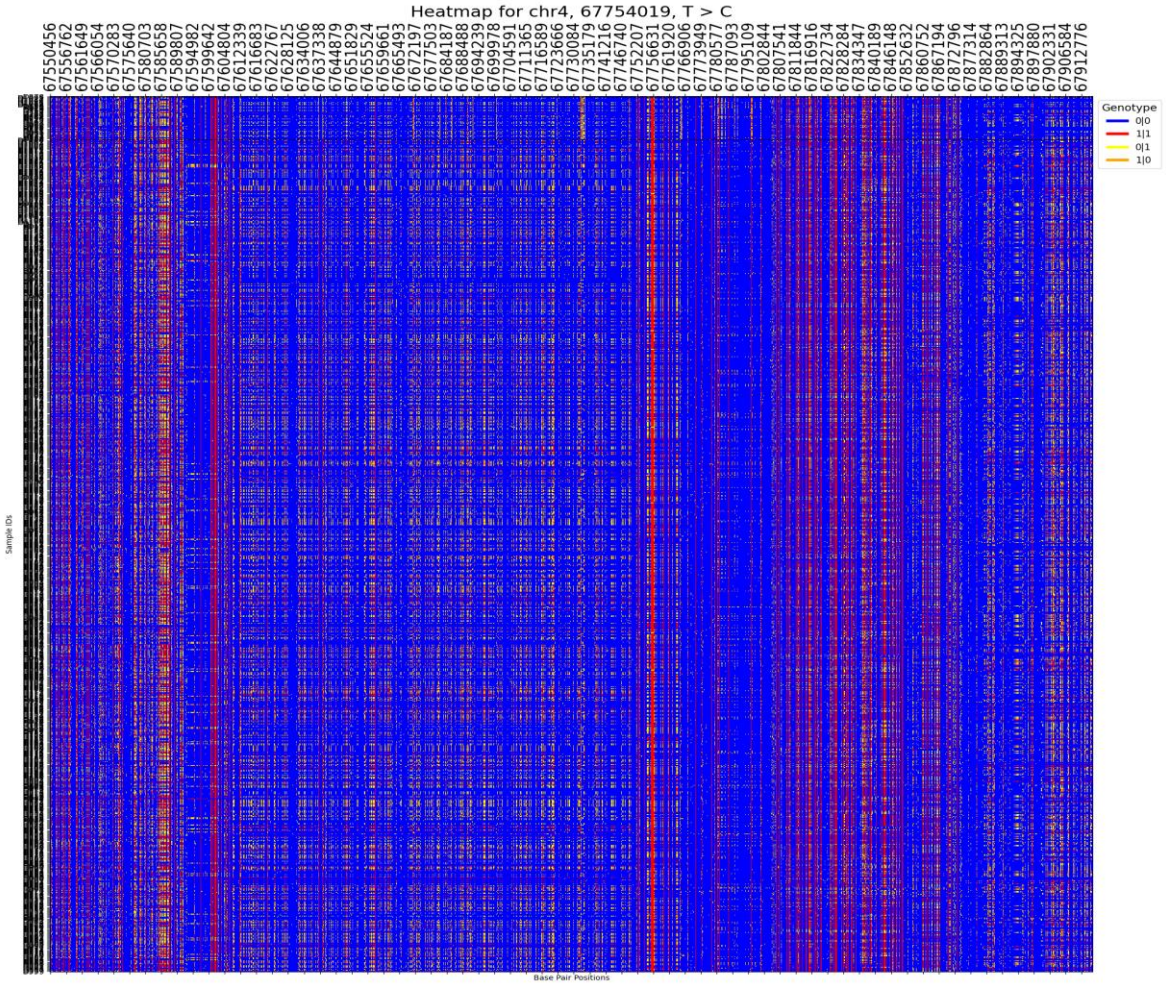

Figure B35: Full heatmap showing the common IBD segment positions for the chromosome 4 variant GNRHR p.Q106R.

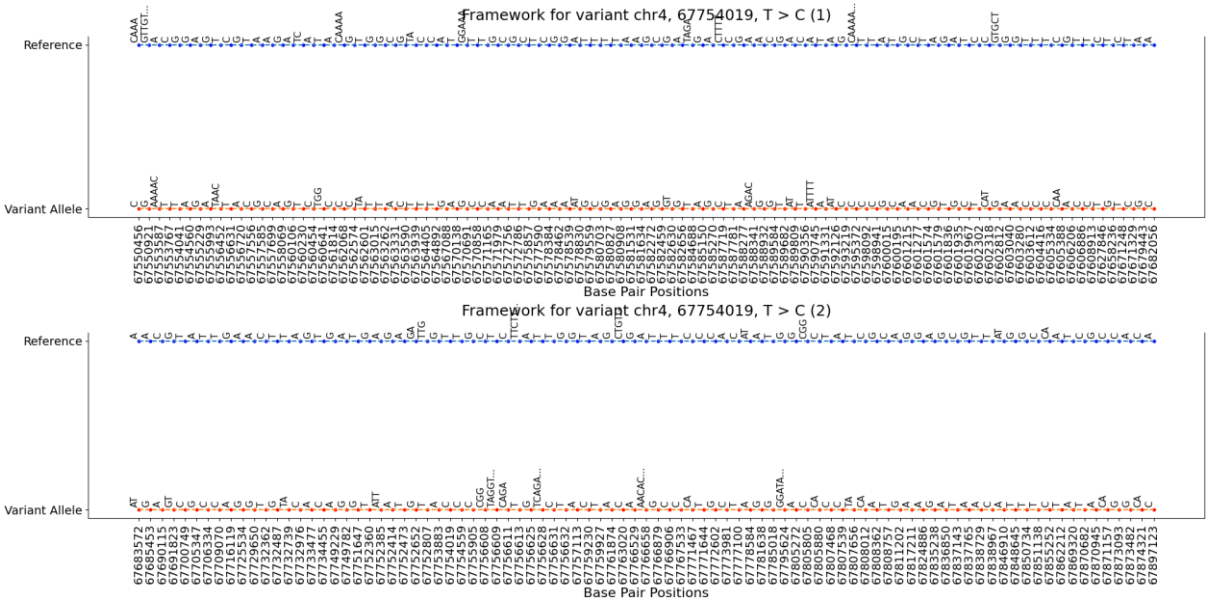

Figure B36: The genetic framework for the individuals that carry the chromosome 4 variant GNRHR p.Q106R

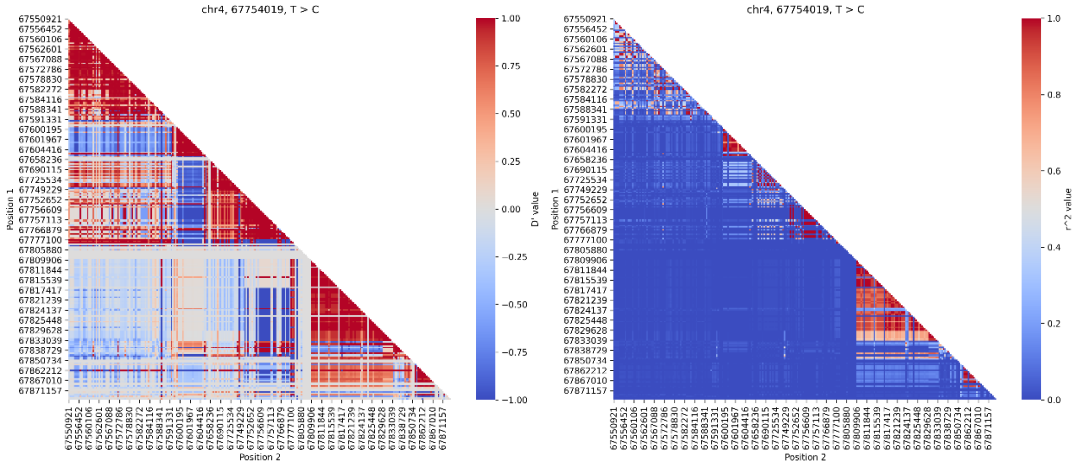

Figure B37: LD plots for the variant framework of the chromosome 4 variant GNRHR p.Q106R, showing high LD in some areas and low LD in others.

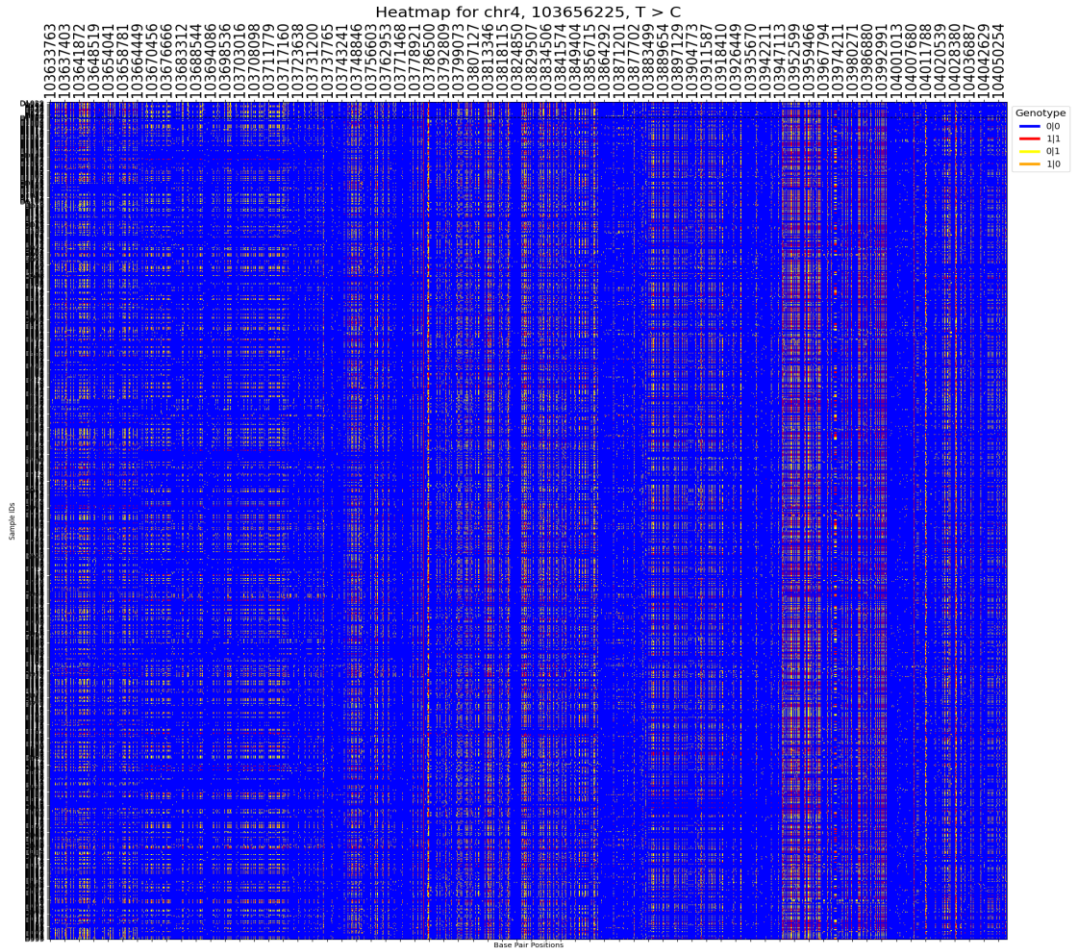

Figure B38: Full heatmap showing the common IBD segment positions for the chromosome 4 variant TACR3 p.K286R.

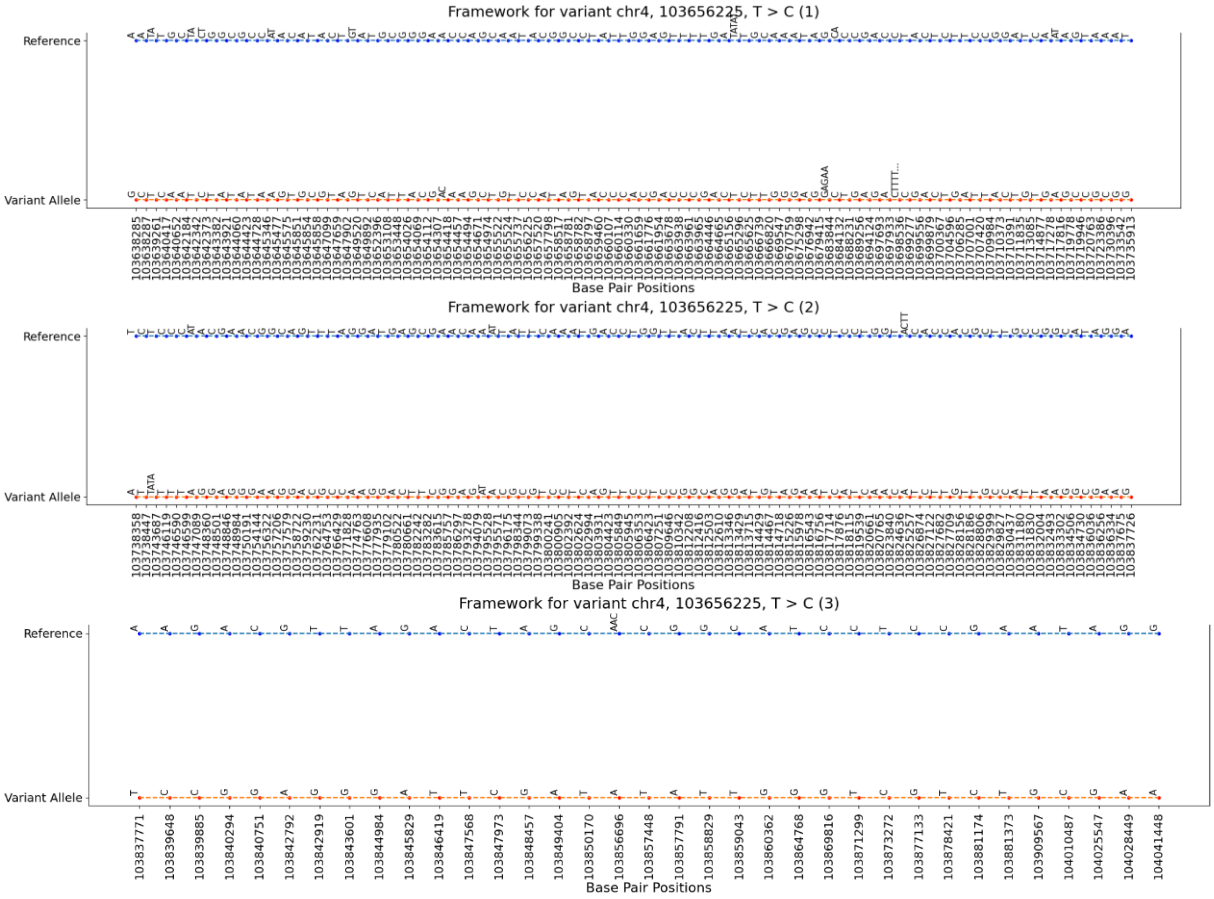

Figure B39: The genetic framework for the individuals that carry the chromosome 4 variant TACR3 p.K286R.

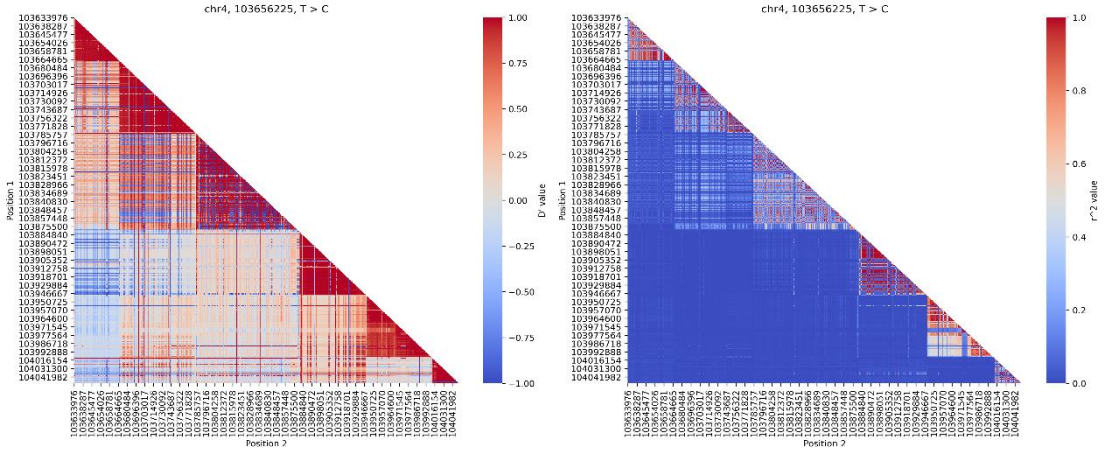

Figure B40: LD plots for the variant framework of the chromosome 4 variant TACR3 p.K286R. The D' plot shows high levels of LD between groups of variants, whereas the r^2 plot shows minimal LD. This small segment of about 0.4cM may have possibly been affected by recombination events over time.

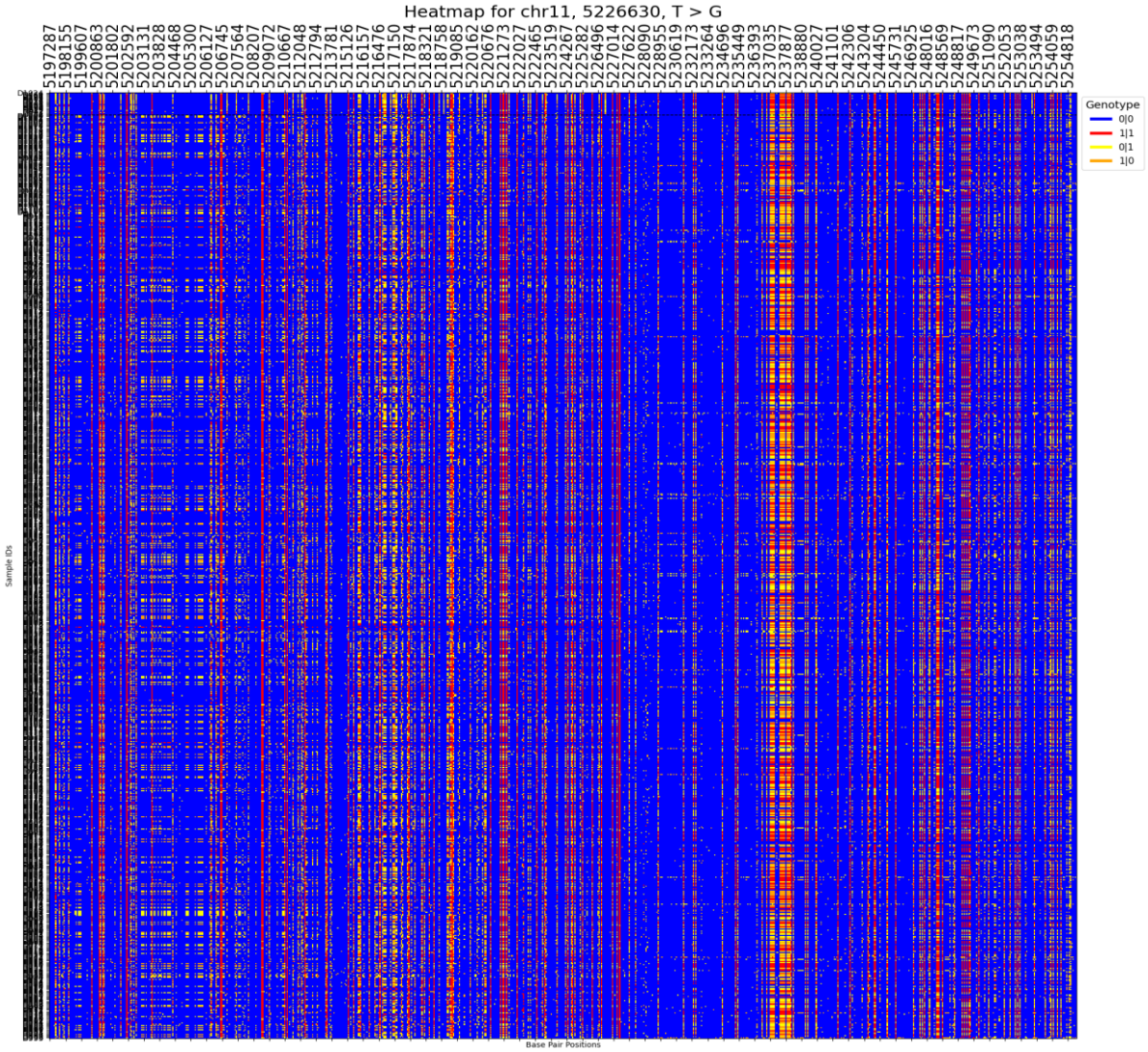

Figure B41: Full heatmap showing the common IBD segment positions for the chromosome 11 variant HBB p.T88P.

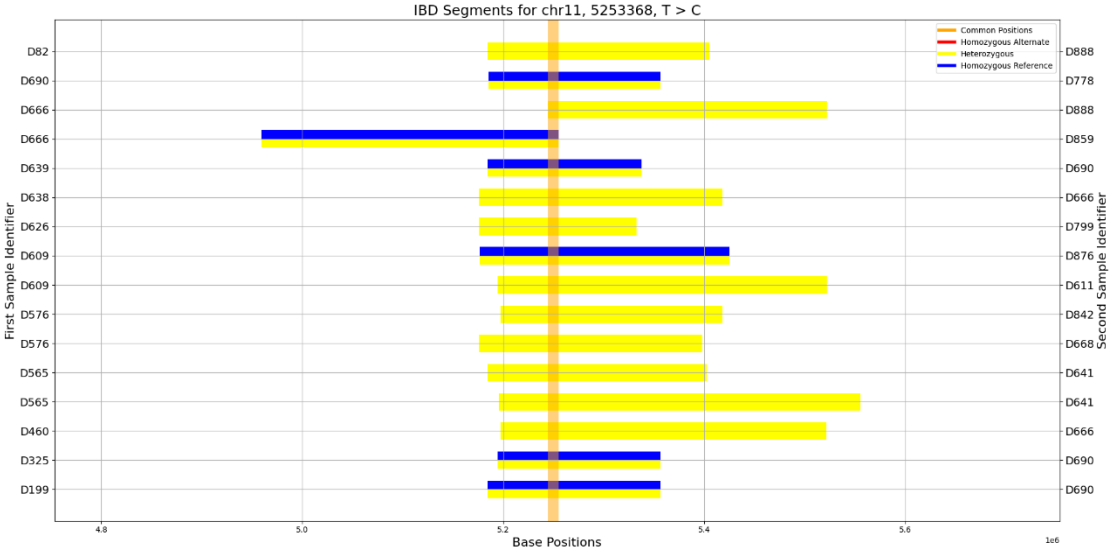

Figure B42: Horizonal bar plot showing the detected IBD segments for chromosome 11 position 5,253,368.

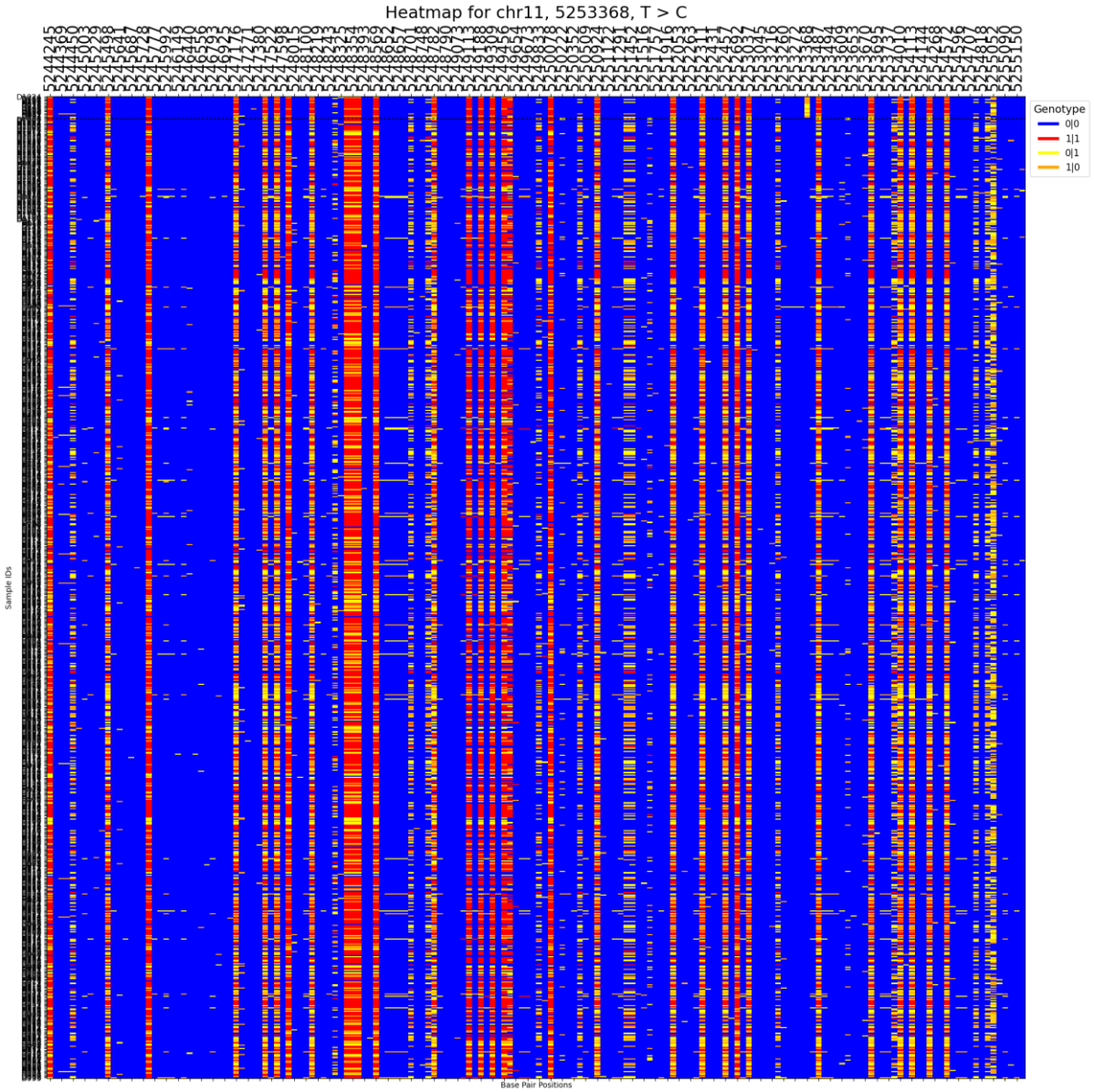

Figure B43: Full heatmap showing the common IBD segment positions for the chromosome 11 variant HBG2 p.H118R.

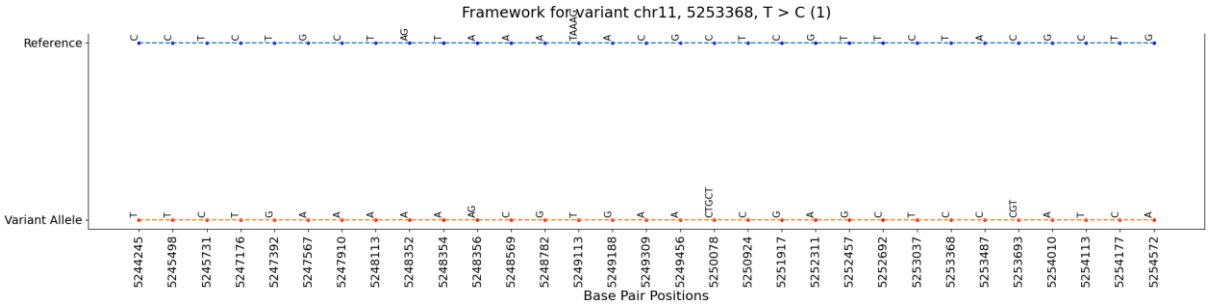

Figure B44: The genetic framework for the individuals that carry the chromosome 11 variant HBG2 p.H118R. This is smaller than the framework represented in HBB p.T88P as the highlighted common position in the horizontal bar plot is smaller and does not include the HBB gene.

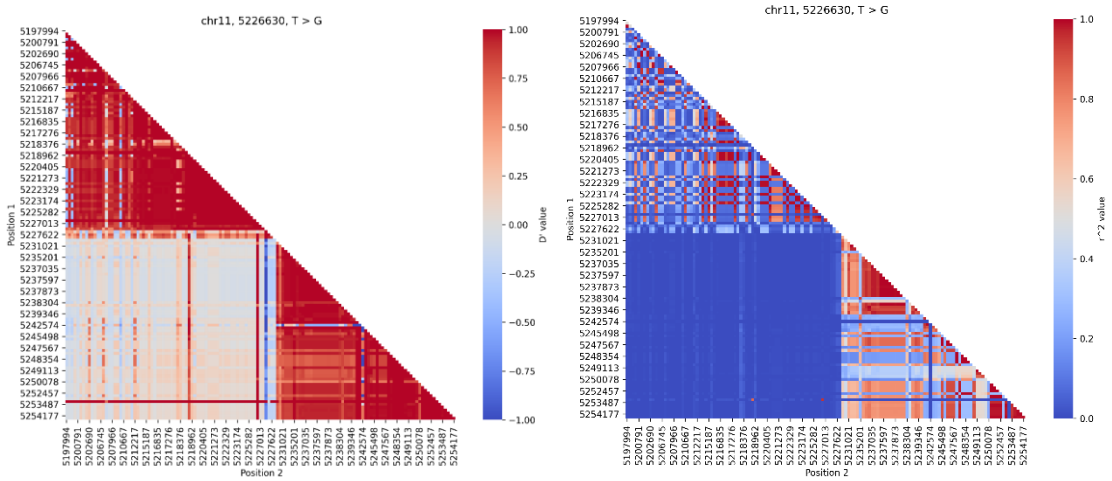

Figure B45: LD plots for the variant framework of the chromosome 11 variants HBB p.T88P and HBG2 p.H118R. The variants of each gene are in LD with the other variants in the same gene. The variants HBB p.T88P and HBG2 p.H118R are in strong LD with a r^2 value of 0.94 and a D' value of 1.

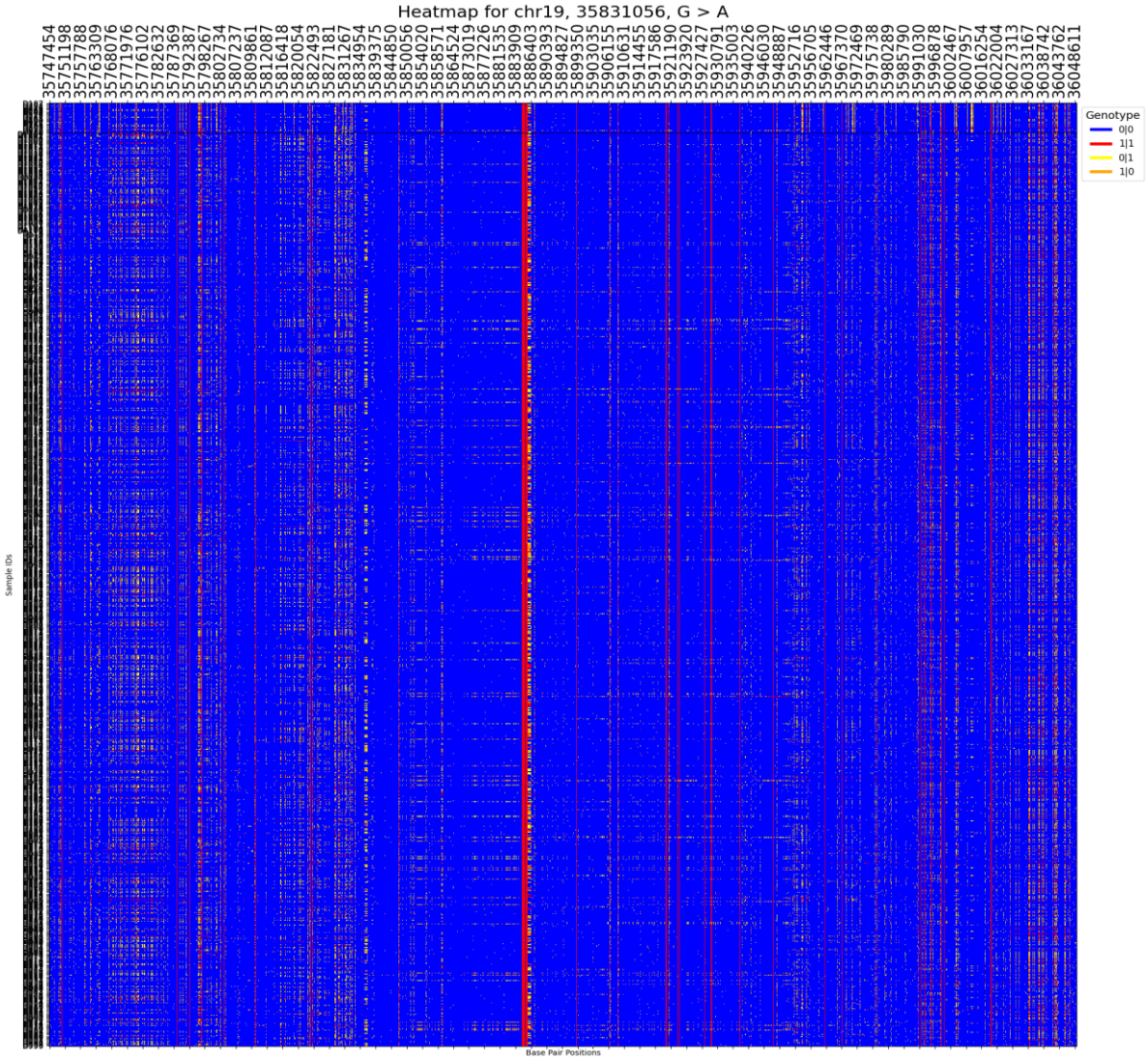

Figure B46: Full heatmap showing the common IBD segment positions for the chromosome 19 variant NPHS1 p.R1160X.

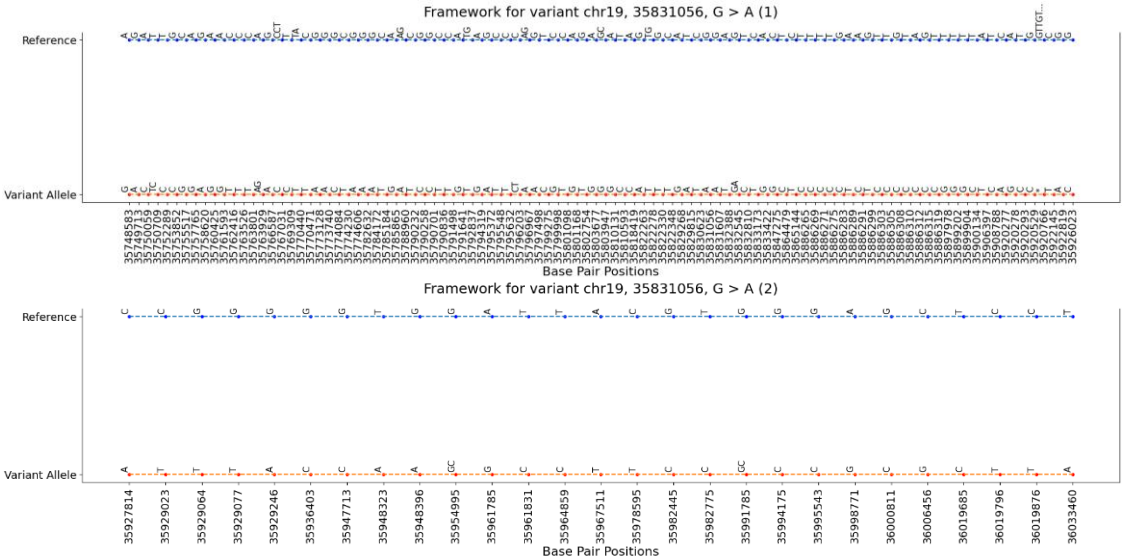

Figure B47: The genetic framework for the individuals that carry the chromosome 19 variant NPHS1 p.R1160X.

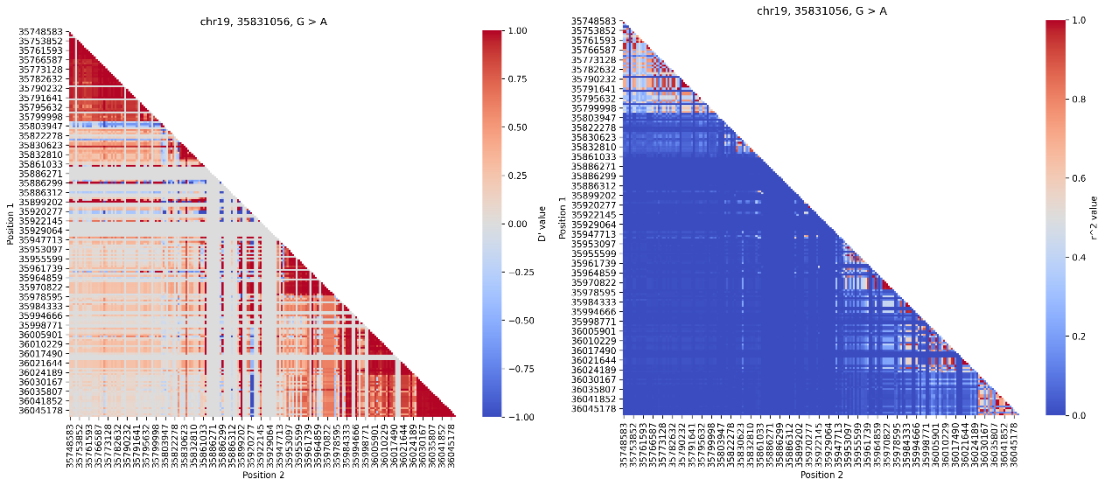

Figure B48: LD plots for the variant framework of the chromosome 19 variant NPHS1 p.R1160X. While some areas show high LD with D', the r^2 LD values show low LD throughout.

Appendix C

The following page includes the media content folder structure for this research project. The contents of the folder are divided into the following directories:

Code: This folder contains all of the code that was developed and used throughout the project. The Python script *variant_search.py* contains the bioinformatics pipeline that was used to generate horizontal bar plots, heatmaps and frameworks. The bash script *generate_ibds.sh* was used to automatically perform dataset filtering, genotype phasing and generation of IBD segments, the latter using the IBD detection tool RaPID with optimised parameters. The Python scripts *plots.py*, *rapid_windows.py* and *rapid_successes.py* were used to generate IBD Benchmark plots to compare the performance of the six IBD detection tools that were tested. The Python scripts *refinedibd.py*, *ibdseq.py* and *query.py* were used to modify the output IBD segment file of the respective tool to match IBD Benchmark's parser. The *README.txt* file contains installation and usage instructions, including all of the dependencies required by these scripts.

Data: This folder contains all of the outputs that were generated through the use of the previously mentioned code, and is further subdivided as follows:

- **ibd_benchmark_outputs:** This subfolder contains the performance results of IBD Benchmark for the six IBD detection tools that were tested. This includes all the outputs that were generated from IBD Benchmark and the plots that were created to visualise the results. RaPID's folder also includes the optimisation results of the tool for different window size and number of successes parameters.
- **variant_ibds:** This subfolder contains the outputs that were generated with the script variant_search.py. This includes horizontal bar plots, heatmaps and genetic frameworks in relation to the variants that were investigated in the project, which are further subdivided into 0.5cM_ibds and 2cM_ibds. The cM values correspond to the minimum length (-d) parameter which was set on RaPID. The ibd_files subfolder contains all of the IBD segments that were detected by RaPID at the 0.5cM and 2cM minimum lengths for each chromosome.

```
202411_MMB5010_DanielCamilleri_codeanddata
├── Code
│   ├── generate_ibds.sh
│   ├── ibdseq.py
│   ├── plots.py
│   ├── query.py
│   ├── rapid_successes.py
│   ├── rapid_windows.py
│   ├── README.txt
│   ├── refinedibd.py
│   └── variant_search.py
└── Data
    ├── ibd_benchmark_outputs
    │   ├── fastsmc_outputs
    │   │   ├── fastsmc_0.01_error
    │   │   ├── fastsmc_0.1_error
    │   │   └── fastsmc_0_error
    │   ├── hapibd_outputs
    │   │   ├── hapibd_0.01_error
    │   │   ├── hapibd_0.1_error
    │   │   └── hapibd_0_error
    │   ├── IBDSeq_outputs
    │   │   ├── genetic_map_GRCh37_chr20.txt
    │   │   ├── ibdseq_0.01_error
    │   │   ├── ibdseq_0.1_error
    │   │   ├── ibdseq_0_error
    │   │   └── ibdseq.py
    │   ├── Plots
    │   │   ├── 0.01_error
    │   │   ├── 0.1_error
    │   │   ├── 0_error
    │   │   ├── plots.py
    │   │   └── rapid_windows
    │   │       ├── 0.01_error
    │   │       ├── 0.1_error
    │   │       └── 0_error
    │   ├── query_outputs
    │   │   ├── query_0.01_error
    │   │   ├── query_0.1_error
    │   │   ├── query_0_error
    │   │   └── query.py
    │   ├── RaPID_outputs
    │   │   ├── rapid_0.01_error
    │   │   ├── rapid_0.1_error
    │   │   ├── rapid_0_error
    │   │   ├── rapid_successes
    │   │   │   ├── 0.01_error
    │   │   │   ├── 0.1_error
    │   │   │   └── rapid_successes.py
    │   │   └── rapid_windows
    │   │       ├── 1
    │   │       │   ├── rapid_0.01_error
    │   │       │   ├── rapid_0.1_error
    │   │       │   └── rapid_0_error
    │   │       ├── 3
    │   │       │   ├── rapid_0.01_error
    │   │       │   ├── rapid_0.1_error
    │   │       │   └── rapid_0_error
    │   │       ├── 30
    │   │       │   ├── rapid_0.01_error
    │   │       │   ├── rapid_0.1_error
    │   │       │   └── rapid_0_error
    │   │       ├── 5
    │   │       │   ├── rapid_0.01_error
    │   │       │   ├── rapid_0.1_error
    │   │       │   └── rapid_0_error
    │   │       ├── fastsmc
    │   │       │   ├── fastsmc_0.01_error
    │   │       │   ├── fastsmc_0.1_error
    │   │       │   └── fastsmc_0_error
    │   │       └── rapid_windows.py
    │   └── refinedibd_outputs
    │       ├── refinedibd_0.01_error
    │       ├── refinedibd_0.1_error
    │       ├── refinedibd_0_error
    │       └── refinedibd.py
    ├── README.txt
    └── variant_ibds
        ├── 0.5cM_ibds
        │   ├── chr11_5226630_T_G
        │   ├── chr11_5253368_T_C
        │   ├── chr1_204190483_CT_C
        │   ├── chr1_204190659_G_C
        │   ├── chr1_204190794_T_C
        │   ├── chr1_54139647_G_GT
        │   ├── chr19_35831056_G_A
        │   ├── chr2_72891345_A_G
        │   ├── chr4_103656225_T_C
        │   └── chr4_67754019_T_C
        ├── 2cM_ibds
        │   ├── chr1_204190483_CT_C
        │   └── chr1_204190659_G_C
        ├── ibd_files
        └── variants.txt
```

References

Adrion, J.R., Cole, C.B., Dukler, N., Galloway, J.G., Gladstein, A.L., Gower, G., Kyriazis, C.C., Ragsdale, A.P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R.A., Durvasula, A., Gronau, I., Kim, B.Y., McKenzie, P., Messer, P.W., Noskova, E., Ortega-Del Vecchyo, D., Racimo, F., Struck, T.J., Gravel, S., Gutenkunst, R.N., Lohmueller, K.E., Ralph, P.L., Schrider, D.R., Siepel, A., Kelleher, J., Kern, A.D., 2020. A community-maintained standard library of population genetic models. eLife 9, e54967. https://doi.org/10.7554/eLife.54967

Ahvenainen, E.K., Hallman, N., Hjelt, L., 1956. Nephrotic syndrome in newborn and young infants. Ann. Paediatr. Fenn. 2, 227–241.

Albalawi, F.S., Daghestani, Maha H., Daghestani, Mazin H., Eldali, A., Warsy, A.S., 2018. rs4889 polymorphism in KISS1 gene, its effect on polycystic ovary syndrome development and anthropometric and hormonal parameters in Saudi women. J. Biomed. Sci. 25, 50. https://doi.org/10.1186/s12929-018-0452-2

Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F.C., Nielsen, R., 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet. Epidemiol. 33, 266–274. https://doi.org/10.1002/gepi.20378

Almeida, J.L., Korch, C.T., 2004. Authentication of Human and Mouse Cell Lines by Short Tandem Repeat (STR) DNA Genotype Analysis, in: Markossian, S., Grossman, A., Arkin, M., Auld, D., Austin, C., Baell, J., Brimacombe, K., Chung, T.D.Y., Coussens, N.P., Dahlin, J.L., Devanarayan, V., Foley, T.L., Glicksman, M., Gorshkov, K., Haas, J.V., Hall, M.D., Hoare, S., Inglese, J., Iversen, P.W., Lal-Nag, M., Li, Z., Manro, J.R., McGee, J., McManus, O., Pearson, M., Riss, T., Saradjian, P., Sittampalam, G.S., Tarselli, M., Trask, O.J., Weidner, J.R., Wildey, M.J., Wilson, K., Xia, M., Xu, X. (Eds.), Assay Guidance Manual. Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD).

Altay, G., Garver, F., Bannister, W.H., Grech, J.L., Felice, A., Huisman, T.H., 1977. Detection and quantitation of the fetal hemoglobin variant Hb F-Malta-I in adults. Biochem. Genet. 15, 915–923. https://doi.org/10.1007/bf00483988

Andrés, A.M., Nowick, K., 2014. Editorial overview: Genetics of human evolution: The genetics of human origins. Curr. Opin. Genet. Dev. 29, v–vii. https://doi.org/10.1016/j.gde.2014.11.001

Attard, R., Dingli, P., Cassar, K., Vassallo, J., Doggen, C., Farrugia, R., Bezzina Wettinger, S., 2014. Conventional risk factors for myocardial infarction in the Maltese population : results from the Maltese Acute Myocardial Infarction (MAMI) study.

Axiak, C.J., Pleven, A., Attard, R., Borg Carbott, F., Ebejer, J.-P., Brincat, I., Cassar, K., Gruppetta, M., Vassallo, J., Bezzina Wettinger, S., Farrugia, R., 2023. High Population Frequency of GNRHR p.Q106R in Malta: An Evaluation of Fertility and Hormone Profiles in Heterozygotes. J. Endocr. Soc. 8, bvad172. https://doi.org/10.1210/jendso/bvad172

Bjelland, D.W., Lingala, U., Patel, P.S., Jones, M., Keller, M.C., 2017. A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data. Eur. J. Hum. Genet. 25, 617–624. https://doi.org/10.1038/ejhg.2017.6

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., Taberlet, P., 2004. How to track and assess genotyping errors in population genetics studies. Mol. Ecol. 13, 3261–3273. https://doi.org/10.1111/j.1365-294X.2004.02346.x

Borg, J., Papadopoulos, P., Georgitsi, M., Gutiérrez, L., Grech, G., Fanis, P., Phylactides, M., Verkerk, A.J.M.H., van der Spek, P.J., Scerri, C.A., Cassar, W., Galdies, R., van Ijcken, W., Ozgür, Z.,

Gillemans, N., Hou, J., Bugeja, M., Grosveld, F.G., von Lindern, M., Felice, A.E., Patrinos, G.P., Philipsen, S., 2010. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. Nat. Genet. 42, 801–805. https://doi.org/10.1038/ng.630

Bowden, S.J., Bodinier, B., Kalliala, I., Zuber, V., Vuckovic, D., Doulgeraki, T., Whitaker, M.D., Wielscher, M., Cartwright, R., Tsilidis, K.K., Bennett, P., Jarvelin, M.-R., Flanagan, J.M., Chadeau-Hyam, M., Kyrgiou, M., FinnGen consortium, 2021. Genetic variation in cervical preinvasive and invasive disease: a genome-wide association study. Lancet Oncol. 22, 548–557. https://doi.org/10.1016/S1470-2045(21)00028-0

Brown, M.D., Glazner, C.G., Zheng, C., Thompson, E.A., 2012. Inferring Coancestry in Population Samples in the Presence of Linkage Disequilibrium. Genetics 190, 1447–1460. https://doi.org/10.1534/genetics.111.137570

Browning, Brian L., Browning, S.R., 2013. Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. 93, 840–851. https://doi.org/10.1016/j.ajhg.2013.09.014

Browning, Brian L, Browning, S.R., 2013. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics 194, 459–471. https://doi.org/10.1534/genetics.113.150029

Browning, B.L., Browning, S.R., 2011. A Fast, Powerful Method for Detecting Identity by Descent. Am. J. Hum. Genet. 88, 173–182. https://doi.org/10.1016/j.ajhg.2011.01.010

Browning, B.L., Browning, S.R., 2007. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet. Epidemiol. 31, 365–375. https://doi.org/10.1002/gepi.20216

Browning, B.L., Tian, X., Zhou, Y., Browning, S.R., 2021. Fast two-stage phasing of large-scale sequence data. Am. J. Hum. Genet. 108, 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Browning, S.R., Browning, B.L., 2010. High-Resolution Detection of Identity by Descent in Unrelated Individuals. Am. J. Hum. Genet. 86, 526–539. https://doi.org/10.1016/j.ajhg.2010.02.021

Camilleri, G., Bezzina Wettinger, S., Farrugia, R., Camilleri, S., 2015. A novel mutation in LRRK2 influences risk for Parkinson disease in the Maltese population.

Capelli, C., Redhead, N., Romano, V., Calì, F., Lefranc, G., Delague, V., Megarbane, A., Felice, A.E., Pascali, V.L., Neophytou, P.I., Poulli, Z., Novelletto, A., Malaspina, P., Terrenato, L., Berebbi, A., Fellous, M., Thomas, M.G., Goldstein, D.B., 2006. Population structure in the Mediterranean basin: a Y chromosome perspective. Ann. Hum. Genet. 70, 207–225. https://doi.org/10.1111/j.1529-8817.2005.00224.x

Cauchi, M.N., Clegg, J.B., Weatherall, D.J., 1969. Haemoglobin F(Malta): a new foetal haemoglobin variant with a high incidence in Maltese infants. Nature 223, 311–313. https://doi.org/10.1038/223311a0

Cavalli-Sforza, L., 1979. The Ashkenazi gene pool: interpretations, in: Goodman, R.M. (Ed.), Genetic Disorders among the Jewish People. Raven Press, New York, pp. 99–104.

Chapelle, A. de la, 1993. Disease gene mapping in isolated human populations: the example of Finland. J. Med. Genet. 30, 857–865. https://doi.org/10.1136/jmg.30.10.857

Charoute, H., Bakhchane, A., Benrahma, H., Romdhane, L., Gabi, K., Rouba, H., Fakiri, M., Abdelhak, S., Lenaers, G., Barakat, A., 2015. Mediterranean Founder Mutation Database (MFMD): Taking

Advantage from Founder Mutations in Genetics Diagnosis, Genetic Diversity and Migration History of the Mediterranean Population. Hum. Mutat. 36, E2441-2453. https://doi.org/10.1002/humu.22835

Chevrier, L., Guimiot, F., de Roux, N., 2011. GnRH receptor mutations in isolated gonadotropic deficiency. Mol. Cell. Endocrinol. 346, 21–28. https://doi.org/10.1016/j.mce.2011.04.018

Collins, A., Morton, N.E., 1998. Mapping a disease locus by allelic association. Proc. Natl. Acad. Sci. 95, 1741–1745. https://doi.org/10.1073/pnas.95.4.1741

Daghestani, Maha H., Daghestani, Mazin H., Daghistani, M., Ambreen, K., Albalawi, F.S., AlNeghery, L.M., Warsy, A.S., 2020. Influence of KISS1 gene polymorphisms on the risk of polycystic ovary syndrome and its associated variables, in Saudi women. BMC Endocr. Disord. 20, 59. https://doi.org/10.1186/s12902-020-0537-2

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., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008

de Roux, N., 2006. GnRH receptor and GPR54 inactivation in isolated gonadotropic deficiency. Best Pract. Res. Clin. Endocrinol. Metab. 20, 515–528. https://doi.org/10.1016/j.beem.2006.10.005

Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., Dermitzakis, E.T., 2019. Accurate, scalable and integrative haplotype estimation. Nat. Commun. 10, 5436. https://doi.org/10.1038/s41467-019-13225-y

Dianzani, I., Howells, D.W., Ponzone, A., Saleeba, J.A., Smooker, P.M., Cotton, R.G., 1993. Two new mutations in the dihydropteridine reductase gene in patients with tetrahydrobiopterin deficiency. J. Med. Genet. 30, 465–469.

Dimitromanolakis, A., Paterson, A.D., Sun, L., 2019. Fast and Accurate Shared Segment Detection and Relatedness Estimation in Un-phased Genetic Data via TRUFFLE. Am. J. Hum. Genet. 105, 78–88. https://doi.org/10.1016/j.ajhg.2019.05.007

Durand, E.Y., Eriksson, N., McLean, C.Y., 2014. Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. Mol. Biol. Evol. 31, 2212–2222. https://doi.org/10.1093/molbev/msu151

Durbin, R., 2014. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics 30, 1266–1272. https://doi.org/10.1093/bioinformatics/btu014

Ekstein, J., Katzenstein, H., 2001. The Dor Yeshorim story: community-based carrier screening for Tay-Sachs disease. Adv. Genet. 44, 297–310. https://doi.org/10.1016/s0065-2660(01)44087-9

Fan, H., Chu, J.-Y., 2007. A Brief Review of Short Tandem Repeat Mutation. Genomics Proteomics Bioinformatics 5, 7–14. https://doi.org/10.1016/S1672-0229(07)60009-6

Fang, H., Wu, Y., Narzisi, G., O'Rawe, J.A., Barrón, L.T.J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M.C., Lyon, G.J., 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 6, 89. https://doi.org/10.1186/s13073-014-0089-z

Farrugia, R., Scerri, C.A., Montalto, S.A., Parascandolo, R., Neville, B.G.R., Felice, A.E., 2007. Molecular genetics of tetrahydrobiopterin (BH4) deficiency in the Maltese population. Mol. Genet. Metab. 90, 277–283. https://doi.org/10.1016/j.ymgme.2006.10.013

Farsimadan, M., Moammadzadeh Ghosi, F., Takamoli, S., Vaziri, H., 2021. Association analysis of KISS1 polymorphisms and haplotypes with polycystic ovary syndrome. Br. J. Biomed. Sci. 78, 201–205. https://doi.org/10.1080/09674845.2020.1864109

Fiorini, S., Mallia-Milanes, V., 1991. Malta: A Case Study in International Cross-currents : Proceedings of the First International Colloquium on the History of the Central Mediterranean Held at the University of Malta, 13-17 December 1989. Malta University Publications.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T.D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, Hongbin, Zhao, Hui, Zhou, J., Gabriel, S.B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R.C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Yan, Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Yayun, Sun, W., Wang, Haifeng, Wang, Yi, Wang, Ying, Xiong, X., Xu, L., Waye, M.M.Y., Tsui, S.K.W., Xue, H., Wong, J.T.-F., Galver, L.M., Fan, J.-B., Gunderson, K., Murray, S.S., Oliphant, A.R., Chee, M.S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M.S., Roumy, S., Sallée, C., Verner, A., Hudson, T.J., Kwok, P.-Y., Cai, D., Koboldt, D.C., Miller, R.D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Qiang Song, Y., Tam, P.K.H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C.P., Delgado, M., Dermitzakis, E.T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B.E., Whittaker, P., Bentley, D.R., Daly, M.J., de Bakker, P.I.W., Barrett, J., Chretien, Y.R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D.J., Sabeti, P., Saxena, R., Schaffner, S.F., Sham, P.C., Varilly, P., Altshuler, D., Stein, L.D., Krishnan, L., Vernon Smith, A., Tello-Ruiz, M.K., Thorisson, G.A., Chakravarti, A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Lin, S., Abecasis, G.R., Guan, W., Li, Y., Munro, H.M., Steve Qin, Z., Thomas, D.J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L.R., Clarke, G., Evans, D.M., Morris, A.P., Weir, B.S., Tsunoda, T., Johnson, T., Mullikin, J.C., Sherry, S.T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, Hui, Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C.N., Adebamowo, C.A., Ajayi, I., Aniagwu, T., Marshall, P.A., Nkwodimmah, C., Royal, C.D.M., Leppert, M.F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I.F., Knoppers, B.M., Foster, M.W., Wright Clayton, E., Watkin, J., Gibbs, R.A., Belmont, J.W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G.M., Wheeler, D.A., Yakub, I., Gabriel, S.B., Onofrio, R.C., Richter, D.J., Ziaugra, L., Birren, B.W., Daly, M.J., Altshuler, D., Wilson, R.K., Fulton, L.L., Rogers, J., Burton, J., Carter, N.P., Clee, C.M., Griffiths, M., Jones, M.C., McLay, K., Plumb, R.W., Ross, M.T., Sims, S.K., Willey, D.L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J.C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, Hongguang, An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A.L., Brooks, L.D., McEwen, J.E., Guyer, M.S., Ota Wang, V., Peterson, J.L., Shi, M., Spiegel, J., Sung, L.M., Zacharia, L.F., Collins, F.S., Kennedy, K., Jamieson, R., The International HapMap Consortium, Genotyping centres: Perlegen Sciences, Baylor College of Medicine and ParAllele BioScience, Beijing Genomics Institute, Broad Institute of Harvard and Massachusetts Institute of Technology, Chinese National Human Genome Center at Beijing, Chinese National Human Genome Center at Shanghai, Chinese University of Hong Kong, Hong Kong University of Science and Technology, Illumina, McGill University and Génome Québec Innovation Centre, University of California at San Francisco and Washington University, University of Hong Kong, University of Tokyo and RIKEN, Wellcome Trust Sanger Institute, Analysis groups: Broad Institute, Cold Spring Harbor Laboratory, Johns Hopkins University School of Medicine, University of Michigan, University of Oxford, University of Oxford, W.T.C. for H.G., RIKEN, US National Institutes of Health, US National Institutes of Health National Center for Biotechnology Information, Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute, Health Sciences University of Hokkaido, E.E.I., and Shinshu

University, Howard University and University of Ibadan, University of Utah, Ethical, legal and social issues: C.A. of S.S., Genetic Interest Group, Kyoto University, Nagasaki University, University of Ibadan School of Medicine, University of Montréal, University of Oklahoma, Vanderbilt University, Wellcome Trust, SNP discovery: Baylor College of Medicine, Washington University, Scientific management: Chinese Academy of Sciences, Genome Canada, Génome Québec, Japanese Ministry of Education, C., Sports, Science and Technology, Ministry of Science and Technology of the People's Republic of China, The Human Genetic Resource Administration of China, The SNP Consortium, 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861. https://doi.org/10.1038/nature06258

Freyman, W.A., McManus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., 23 and Me Research Team, Auton, A., 2021. Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows-Wheeler Transform. Mol. Biol. Evol. 38, 2131–2151. https://doi.org/10.1093/molbev/msaa328

Gianetti, E., Hall, J.E., Au, M.G., Kaiser, U.B., Quinton, R., Stewart, J.A., Metzger, D.L., Pitteloud, N., Mericq, V., Merino, P.M., Levitsky, L.L., Izatt, L., Lang-Muritano, M., Fujimoto, V.Y., Dluhy, R.G., Chase, M.L., Crowley, W.F., Plummer, L., Seminara, S.B., 2012. When genetic load does not correlate with phenotypic spectrum: lessons from the GnRH receptor (GNRHR). J. Clin. Endocrinol. Metab. 97, E1798-1807. https://doi.org/10.1210/jc.2012-1264

Gong, B., Lababidi, S., Kusko, R., Bouri, K., Prezek, S., Thovarai, V., Prasanna, A., Maier, E.J., Golkaram, M., Sun, X., Kyriakidis, K., Kitajima, J.P., Ebrahim Sahraeian, S.M., Guo, Y., Johanson, E., Jones, W., Tong, W., Xu, J., 2024. Towards accurate indel calling for oncopanel sequencing through an international pipeline competition at precisionFDA. Sci. Rep. 14, 8165. https://doi.org/10.1038/s41598-024-58573-y

Greene, D.N., Vaughn, C.P., Crews, B.O., Agarwal, A.M., 2015. Advances in detection of hemoglobinopathies. Clin. Chim. Acta 439, 50–57. https://doi.org/10.1016/j.cca.2014.10.006

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., Pe'er, I., 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19, 318–326. https://doi.org/10.1101/gr.081398.108

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet. 5, e1000695. https://doi.org/10.1371/journal.pgen.1000695

Han, L., Abney, M., 2011. Identity by Descent Estimation With Dense Genome-Wide Genotype Data. Genet. Epidemiol. 35, 557–567. https://doi.org/10.1002/gepi.20606

Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., Lander, E., 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. 2, 204–211. https://doi.org/10.1038/ng1192-204

Henden, L., Lee, S., Mueller, I., Barry, A., Bahlo, M., 2018. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genet. 14, e1007279. https://doi.org/10.1371/journal.pgen.1007279

Heshusius, S., Grech, L., Gillemans, N., Brouwer, R.W.W., den Dekker, X.T., van IJcken, W.F.J., Nota, B., Felice, A.E., van Dijk, T.B., von Lindern, M., Borg, J., van den Akker, E., Philipsen, S., 2022. Epigenomic analysis of KLF1 haploinsufficiency in primary human erythroblasts. Sci. Rep. 12, 336. https://doi.org/10.1038/s41598-021-04126-6

Hochreiter, S., 2013. HapFABIA: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res. 41, e202. https://doi.org/10.1093/nar/gkt1013

Howard, S.R., 2019. The Genetic Basis of Delayed Puberty. Front. Endocrinol. 10, 423. https://doi.org/10.3389/fendo.2019.00423

Huttunen, N.P., 1976. Congenital nephrotic syndrome of Finnish type. Study of 75 patients. Arch. Dis. Child. 51, 344–348. https://doi.org/10.1136/adc.51.5.344

Jain, A., Sharma, D., Bajaj, A., Gupta, V., Scaria, V., 2021. Founder variants and population genomes—Toward precision medicine, in: Advances in Genetics. Elsevier, pp. 121–152. https://doi.org/10.1016/bs.adgen.2020.11.004

Jardón-Valadez, E., Ulloa-Aguirre, A., Piñeiro, A., 2008. Modeling and molecular dynamics simulation of the human gonadotropin-releasing hormone receptor in a lipid bilayer. J. Phys. Chem. B 112, 10704–10713. https://doi.org/10.1021/jp800544x

Kääriäinen, H., Muilu, J., Perola, M., Kristiansson, K., 2017. Genetics in an isolated population like Finland: a different basis for genomic medicine? J. Community Genet. 8, 319–326. https://doi.org/10.1007/s12687-017-0318-4

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., Neale, B.M., Daly, M.J., MacArthur, D.G., 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kelleher, J., Etheridge, A.M., McVean, G., 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput. Biol. 12, e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Khoshnoodi, J., Tryggvason, K., 2001. Congenital nephrotic syndromes. Curr. Opin. Genet. Dev. 11, 322–327. https://doi.org/10.1016/S0959-437X(00)00197-0

Koren, A., Zalman, L., Palmor, H., Ekstein, E., Schneour, Y., Schneour, A., Shalev, S., Rachmilewitz, E.A., Filon, D., Openhaim, A., 2002. [The prevention programs for beta thalassemia in the Jezreel and Eiron valleys: results of fifteen years experience]. Harefuah 141, 938–943, 1210.

Koziell, A., Grech, V., Hussain, S., Lee, G., Lenkkeri, U., Tryggvason, K., Scambler, P., 2002. Genotype/phenotype correlations of NPHS1 and NPHS2 mutations in nephrotic syndrome advocate a functional inter-relationship in glomerular filtration. Hum. Mol. Genet. 11, 379–388. https://doi.org/10.1093/hmg/11.4.379

Krstevska-Konstantinova, M., Jovanovska, J., Tasic, V.B., Montenegro, L.R., Beneduzzi, D., Silveira, L.F.G., Gucev, Z.S., 2014. Mutational analysis of KISS1 and KISS1R in idiopathic central precocious puberty. J. Pediatr. Endocrinol. Metab. 27, 199–201. https://doi.org/10.1515/jpem-2013-0080

Kusters, D.M., Huijgen, R., Defesche, J.C., Vissers, M.N., Kindt, I., Hutten, B.A., Kastelein, J.J.P., 2011. Founder mutations in the Netherlands: geographical distribution of the most prevalent mutations in the low-density lipoprotein receptor and apolipoprotein B genes. Neth. Heart J. 19, 175–182. https://doi.org/10.1007/s12471-011-0076-6

Kutlar, F., Felice, A.E., Grech, J.L., Bannister, W.H., Kutlar, A., Wilson, J.B., Webber, B.B., Hu, H.Y., Huisman, T.H., 1991. The linkage of Hb Valletta [alpha 2 beta 287(f3)Thr----Pro] and Hb F-Malta-

I [alpha 2G gamma 2117(G19)His----Arg] in the Maltese population. Hum. Genet. 86, 591–594. https://doi.org/10.1007/bf00201546

Lasaga, M., Debeljuk, L., 2011. Tachykinins and the hypothalamo–pituitary–gonadal axis: An update. Peptides 32, 1972–1978. https://doi.org/10.1016/j.peptides.2011.07.009

Lenkkeri, U., Männikkö, M., McCready, P., Lamerdin, J., Gribouval, O., Niaudet, P.M., Antignac C, K., Kashtan, C.E., Homberg, C., Olsen, A., Kestilä, M., Tryggvason, K., 1999. Structure of the gene for congenital nephrotic syndrome of the finnish type (NPHS1) and characterization of mutations. Am. J. Hum. Genet. 64, 51–61. https://doi.org/10.1086/302182

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, N., Stephens, M., 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213–2233.

Manchanda, R., Legood, R., Burnell, M., McGuire, A., Raikou, M., Loggenberg, K., Wardle, J., Sanderson, S., Gessler, S., Side, L., Balogun, N., Desai, R., Kumar, A., Dorkins, H., Wallis, Y., Chapman, C., Taylor, R., Jacobs, C., Tomlinson, I., Beller, U., Menon, U., Jacobs, I., 2015. Cost-effectiveness of population screening for BRCA mutations in Ashkenazi jewish women compared with family history-based testing. J. Natl. Cancer Inst. 107, 380. https://doi.org/10.1093/jnci/dju380

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., Chen, W.-M., 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Martins Trevisan, C., Naslavsky, M.S., Monfardini, F., Wang, J., Zatz, M., Peluso, C., Pellegrino, R., Mafra, F., Hakonarson, H., Ferreira, F.M., Nakaya, H., Christofolini, D.M., Montagna, E., Crandall, K.A., Barbosa, C.P., Bianco, B., 2020. Variants in the Kisspeptin-GnRH Pathway Modulate the Hormonal Profile and Reproductive Outcomes. DNA Cell Biol. 39, 1012–1022. https://doi.org/10.1089/dna.2019.5165

Mathijssen, I.B., van Maarle, M.C., Kleiss, I.I.M., Redeker, E.J.W., ten Kate, L.P., Henneman, L., Meijers-Heijboer, H., 2017. With expanded carrier screening, founder populations run the risk of being overlooked. J. Community Genet. 8, 327–333. https://doi.org/10.1007/s12687-017-0309-5

Mejri, A., Siala, H., Ouali, F., Bibi, A., Messaoud, T., 2012. Identification of candidate genes involved in clinical variability among Tunisian patients with β-thalassemia. Gene 506, 166–172. https://doi.org/10.1016/j.gene.2012.06.078

Meng, F., Zhao, A., Lu, H., Zou, D., Dong, B., Wang, X., Liu, L., Zhou, S., 2023. KISS1 Gene Variations and Susceptibility to Idiopathic Recurrent Pregnancy Loss. Reprod. Sci. Thousand Oaks Calif 30, 2573–2579. https://doi.org/10.1007/s43032-023-01203-1

Metcalfe, K.A., Poll, A., Royer, R., Llacuachaqui, M., Tulman, A., Sun, P., Narod, S.A., 2010. Screening for founder mutations in BRCA1 and BRCA2 in unselected Jewish women. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 28, 387–391. https://doi.org/10.1200/JCO.2009.25.0712

Microsoft, 2024. C# | Modern, open-source programming language for .NET [WWW Document]. Microsoft. URL https://dotnet.microsoft.com/en-us/languages/csharp (accessed 4.6.24).

Mono Project, 2024. C# Compiler | Mono [WWW Document]. C Compil. URL https://www.mono-project.com/docs/about-mono/languages/csharp/ (accessed 4.6.24).

Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Giménez, J., Reis, A., Varon-Mateeva, R., Macek, M., Kalaydjieva, L., 1994. The origin of the major cystic fibrosis mutation (delta F508) in European populations. Nat. Genet. 7, 169–175. https://doi.org/10.1038/ng0694-169

Nait Saada, J., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., Palamara, P.F., 2020. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat. Commun. 11, 6130. https://doi.org/10.1038/s41467-020-19588-x

Naseri, A., Liu, X., Tang, K., Zhang, S., Zhi, D., 2019. RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. Genome Biol. 20, 143. https://doi.org/10.1186/s13059-019-1754-8

Nei, M., Maruyama, T., Chakraborty, R., 1975. The Bottleneck Effect and Genetic Variability in Populations. Evolution 29, 1–10. https://doi.org/10.2307/2407137

Neville, B.G.R., Parascandalo, R., Farrugia, R., Felice, A., 2005. Sepiapterin reductase deficiency: a congenital dopa-responsive motor and cognitive disorder. Brain J. Neurol. 128, 2291–2296. https://doi.org/10.1093/brain/awh603

Norio, R., 1966. Heredity in the congenital nephrotic syndrome. A genetic study of 57 finnish FAMILIES WITH A REVIEW OF REPORTED CASES. Ann. Paediatr. Fenn. 12, Suppl 27:1-94.

Nyström-Lahti, M., Sistonen, P., Mecklin, J.P., Pylkkänen, L., Aaltonen, L.A., Järvinen, H., Weissenbach, J., de la Chapelle, A., Peltomäki, P., 1994. Close linkage to chromosome 3p and conservation of ancestral founding haplotype in hereditary nonpolyposis colorectal cancer families. Proc. Natl. Acad. Sci. U. S. A. 91, 6054–6058. https://doi.org/10.1073/pnas.91.13.6054

Palomaki, G.E., 2015. Screening for breast cancer by molecular testing for three founder mutations in the BRCA1 and BRCA2 genes among women of Ashkenazi Jewish heritage. J. Med. Screen. 22, 109–111. https://doi.org/10.1177/0969141315579701

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286. https://doi.org/10.1109/5.18626

Ralph, P., Coop, G., 2013. The Geography of Recent Genetic Ancestry across Europe. PLOS Biol. 11, e1001555. https://doi.org/10.1371/journal.pbio.1001555

Ramakrishnan, A.P., 2013. Linkage Disequilibrium, in: Maloy, S., Hughes, K. (Eds.), Brenner's Encyclopedia of Genetics (Second Edition). Academic Press, San Diego, pp. 252–253. https://doi.org/10.1016/B978-0-12-374984-0.00870-6

Reeve, J.P., Rannala, B., 2002. DMLE+: Bayesian linkage disequilibrium gene mapping. Bioinforma. Oxf. Engl. 18, 894–895. https://doi.org/10.1093/bioinformatics/18.6.894

Roa, B.B., Boyd, A.A., Volcik, K., Richards, C.S., 1996. Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. Nat. Genet. 14, 185–187. https://doi.org/10.1038/ng1096-185

Rodriguez, J.M., Bercovici, S., Huang, L., Frostig, R., Batzoglou, S., 2015. Parente2: a fast and accurate method for detecting identity by descent. Genome Res. 25, 280–289. https://doi.org/10.1101/gr.173641.114

Rossum, G.V., Drake, F.L., 2009. Python 3 Reference Manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform.

Rubinstein, W.S., 2004. Hereditary breast cancer in Jews. Fam. Cancer 3, 249–257. https://doi.org/10.1007/s10689-004-9550-2

Ruitberg, C.M., Reeder, D.J., Butler, J.M., 2001. STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. 29, 320–322.

Said, E., Chong, J.X., Hempel, M., Denecke, J., Soler, P., Strom, T., Nickerson, D.A., Kubisch, C., Bamshad, M., Lessel, D., 2017. Survival Beyond the Perinatal Period Expands the Phenotypes Caused by Mutations in GLE1. Am. J. Med. Genet. A. 173, 3098–3103. https://doi.org/10.1002/ajmg.a.38406

Sanaullah, A., Zhi, D., Zhang, S., 2021. d-PBWT: dynamic positional Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 37, 2390–2397. https://doi.org/10.1093/bioinformatics/btab117

Seidman, D.N., Shenoy, S.A., Kim, M., Babu, R., Woods, I.G., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Williams, A.L., 2020. Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. Am. J. Hum. Genet. 106, 453–466. https://doi.org/10.1016/j.ajhg.2020.02.012

Shemirani, R., Belbin, G.M., Avery, C.L., Kenny, E.E., Gignoux, C.R., Ambite, J.L., 2021. Rapid detection of identity-by-descent tracts for mega-scale datasets. Nat. Commun. 12, 3546. https://doi.org/10.1038/s41467-021-22910-w

Slatkin, M., 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9, 477–485. https://doi.org/10.1038/nrg2361

Sliz, E., Huilaja, L., Pasanen, A., Laisk, T., Reimann, E., Mägi, R., FinnGen, Estonian Biobank Research Team, Hannula-Jouppi, K., Peltonen, S., Salmi, T., Koulu, L., Tasanen, K., Kettunen, J., 2022. Uniting biobank resources reveals novel genetic pathways modulating susceptibility for atopic dermatitis. J. Allergy Clin. Immunol. 149, 1105-1112.e9. https://doi.org/10.1016/j.jaci.2021.07.043

Stephen, S.B., Pauline, R., Velmurugan, S., Subbaraj, G.K., 2024. An association between fat mass and obesity-associated (FTO) (rs9939609) and kisspeptin-1 (KISS-1) (rs4889, rs372790354) gene polymorphisms with polycystic ovary syndrome: an updated meta-analysis and power analysis. J. Assist. Reprod. Genet. 41, 2457–2475. https://doi.org/10.1007/s10815-024-03213-7

Sticca, E.L., Belbin, G.M., Gignoux, C.R., 2021. Current Developments in Detection of Identity-by-Descent Methods and Applications. Front. Genet. 12.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 12, e1001779. https://doi.org/10.1371/journal.pmed.1001779

Tang, K., Naseri, A., Wei, Y., Zhang, S., Zhi, D., 2022. Open-source benchmarking of IBD segment detection methods for biobank-scale cohorts. GigaScience 11, giac111. https://doi.org/10.1093/gigascience/giac111

Tataru, P., Nirody, J.A., Song, Y.S., 2014. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. Bioinformatics 30, 3430–3431. https://doi.org/10.1093/bioinformatics/btu563

Ter Haar, N.M., Jeyaratnam, J., Lachmann, H.J., Simon, A., Brogan, P.A., Doglio, M., Cattalini, M., Anton, J., Modesto, C., Quartier, P., Hoppenreijs, E., Martino, S., Insalaco, A., Cantarini, L., Lepore, L., Alessio, M., Calvo Penades, I., Boros, C., Consolini, R., Rigante, D., Russo, R., Pachlopnik Schmid, J., Lane, T., Martini, A., Ruperto, N., Frenkel, J., Gattorno, M., Paediatric Rheumatology International Trials Organisation and Eurofever Project, 2016. The Phenotype and Genotype of Mevalonate Kinase Deficiency: A Series of 114 Cases From the Eurofever Registry. Arthritis Rheumatol. Hoboken NJ 68, 2795–2805. https://doi.org/10.1002/art.39763

Terwilliger, J.D., 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am. J. Hum. Genet. 56, 777–787.

Thompson, E.A., 2013. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. Genetics 194, 301–326. https://doi.org/10.1534/genetics.112.148825

Thorlacius, S., Sigurdsson, S., Bjarnadottir, H., Olafsdottir, G., Jonasson, J.G., Tryggvadottir, L., Tulinius, H., Eyfjörd, J.E., 1997. Study of a single BRCA2 mutation with high carrier frequency in a small population. Am. J. Hum. Genet. 60, 1079–1084.

Tusset, C., Noel, S.D., Trarbach, E.B., Silveira, L.F.G., Jorge, A.A.L., Brito, V.N., Cukier, P., Seminara, S.B., de Mendonça, B.B., Kaiser, U.B., Latronico, A.C., 2012. Mutational Analysis of TAC3 and TACR3 Genes in Patients with Idiopathic Central Pubertal Disorders. Arq. Bras. Endocrinol. Metabol. 56, 646–652.

Uusimaa, J., Kettunen, J., Varilo, T., Järvelä, I., Kallijärvi, J., Kääriäinen, H., Laine, M., Lapatto, R., Myllynen, P., Niinikoski, H., Rahikkala, E., Suomalainen, A., Tikkanen, R., Tyynismaa, H., Vieira, P., Zarybnicky, T., Sipilä, P., Kuure, S., Hinttala, R., 2022. The Finnish genetic heritage in 2022 – from diagnosis to translational research. Dis. Model. Mech. 15, dmm049490. https://doi.org/10.1242/dmm.049490

Vatsiou, A.I., Bazin, E., Gaggiotti, O.E., 2016. Changes in selective pressures associated with human population expansion may explain metabolic and immune related pathways enriched for signatures of positive selection. BMC Genomics 17, 504. https://doi.org/10.1186/s12864-016-2783-2

Virtaneva, K., Miao, J., Träskelin, A.L., Stone, N., Warrington, J.A., Weissenbach, J., Myers, R.M., Cox, D.R., Sistonen, P., de la Chapelle, A., 1996. Progressive myoclonus epilepsy EPM1 locus maps to a 175-kb interval in distal 21q. Am. J. Hum. Genet. 58, 1247–1253.

Wall, J.D., Tang, L.F., Zerbe, B., Kvale, M.N., Kwok, P.-Y., Schaefer, C., Risch, N., 2014. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res. 24, 1734–1739. https://doi.org/10.1101/gr.168393.113

Wallace, S.E., Bean, L.J., 2021. Resources for Genetics Professionals — Genetic Disorders Associated with Founder Variants Common in the Inuit Population, GeneReviews® [Internet]. University of Washington, Seattle.

Wang, C., Veldsman, W.P., Zhang, L., 2023. Detection of short identity by descent segments using low-frequency variants. https://doi.org/10.1101/2023.09.26.559464

Wei, Y., Naseri, A., Zhi, D., Zhang, S., 2023. RaPID-Query for fast identity by descent search and genealogical analysis. Bioinformatics 39, btad312. https://doi.org/10.1093/bioinformatics/btad312

Wojciechowski, P., Lipowska, A., Rys, P., Ewens, K.G., Franks, S., Tan, S., Lerchbaum, E., Vcelak, J., Attaoua, R., Straczkowski, M., Azziz, R., Barber, T.M., Hinney, A., Obermayer-Pietsch, B., Lukasova, P., Bendlova, B., Grigorescu, F., Kowalska, I., Goodarzi, M.O., Strauss, J.F., McCarthy, M.I., Malecki, M.T., 2012. Impact of FTO genotypes on BMI and weight in polycystic ovary

syndrome: a systematic review and meta-analysis. Diabetologia 55, 2636–2645. https://doi.org/10.1007/s00125-012-2638-6

Xiong, M., Guo, S.-W., 1997. Fine-Scale Genetic Mapping Based on Linkage Disequilibrium: Theory and Applications. Am. J. Hum. Genet. 60, 1513–1531. https://doi.org/10.1086/515475

Zhou, Y., Browning, S.R., Browning, B.L., 2020. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. Am. J. Hum. Genet. 106, 426–437. https://doi.org/10.1016/j.ajhg.2020.02.010

Zhu, N., Zhao, M., Song, Y., Ding, L., Ni, Y., 2022. The KiSS-1/GPR54 system: Essential roles in physiological homeostasis and cancer biology. Genes Dis. 9, 28–40. https://doi.org/10.1016/j.gendis.2020.07.008