Palestine Polytechnic University

Deanship of graduate studies and Scientific Research

Bethlehem University

Faculty of Science

Define Runs of Homozygosity in a Cohort of Palestinian Families

By

Yasmin Khaled Omar Tamimi

This thesis is submitted in partial fulfillment of the requirements for a Master Degree in Science

August, 2016

The undersigned hereby certify that they have read and recommended to the Faculty of Scientific Research and Higher Studies at the Palestine Polytechnic University and the Faculty of Science at Bethlehem University the acceptance the thesis entitled:

Define Runs of Homozygosity in a Cohort of Palestinian Families
By
Yasmin Khaled Omar Tamimi

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Biotechnology

Graduate Advisory Committee:

Committee Member (Student's supervisor) *Prof.Mo'ien Kan'an, Bethlehem University*	Date
Committee Member(Internal Examiner) *Dr.Omar Mousa Darissa, Bethlehem University*	Date
Committee Member (External Examiner) *Dr. Siba Shanak, The Arab American University-Jenin (AAUJ)*	Date

Approved by the Faculties

Dean of Faculty of Scientific Research and Higher studies Palestine Polytechnic University	Dean of Faculty of Science Bethlehem University
Date	Date

Define Runs of Homozygosity in a Cohort of Palestinian Families

By

Yasmin Khaled Omar Tamimi

ABSTRACT

Objective: The aim of this study is to define runs of homozygosity (ROH) in a cohort of Palestinian families.

Methods: Defining ROH depends on using data of single nucleotide polymorphism (SNP) locations, and the lengths of shared homozygous segments. By using data from 197,701 SNP panel in 29 individuals from southern area of Palestine, we reflect the potential significance of defining ROH.

Results: In all individuals, regions of ROH were identified; these regions were mapped and their genes content were revaluated.

Conclusion: Genes located within ROH regions may have significant causes of deleterious variants that could be inherited in concordance with homozygosity runs.

تعريف الامتدادات المتماثلة في مجموعه من الأسر الفلسطينية

ياسمين خالد عمر التميمي

ملخص

تهدف هذه الدراسة الى تعريف تطابق الأليلات (ROH) في الجينوم لدى مجموعه من الأسر الفلسطينية. المنهجية: يعتمد تعريف هذه الامتدادات على استخدام البيانات الخاصة بالـ(SNPs) من حيث موقعها وحجم الامتدادات المتماثلة لها. و باستخدام بيانات من 29 فرد بعدد SNPs 197,701 من منطقه جنوب فلسطين، يمكن أن نعكس ضرورة تعريف الـ ROH. تبين النتائج أن تم التعرف على الـ ROH في كل الافراد على طول الجينوم، وتم تعيين المناطق المرتبطة بحدوث أمراض متصلة بجينات. نستنتج أن الجينات الواقعة في مناطق الـ ROH تحتوي على مسببات هامه للجينات المؤذية.

DECLARATION

I hereby declare that the Master Thesis entitled " Define Runs of Homozygosity in a Cohort of Palestinian families " is my own original work, and hereby certify that unless stated otherwise, all work contained within this thesis is from my own independent research, and it has not been submitted for the award of any other degree at any institution, except where due acknowledgment is made in the text.

Name and Signature: Yasmin Khaled Omar Tamimi

Date: August, 2016

STATEMENT OF PERMISSION TO USE

In presenting this thesis, in partial fulfillment of the requirements for the joint Master degree in Biotechnology at Palestine Polytechnic University and Bethlehem University, I hereby agree that the library shall make it available to borrowers in accordance with library regulations. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of the source is made.

Permission for extensive quotations from, reproduction, or publication of this thesis may be granted by main supervisor, or in [his/her] absence, by Dean of Higher Studies when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without author written permission.

Signature: Yasmin Khaled Omar Tamimi

Date: August, 2016

ACKNOWLEDGMENT

I would first like to thank my thesis advisor Prof. Mo'ien Kanaan, coordinator of the Biotechnology Masters at Bethlehem University. Prof. Kanaan gave me the opportunity to work in the field of bioinformatics, the nearest specialty to my field in computer science.

I would also like to thank Dr. Fuad Zahdeh, PhD student in Bioinformatics at Hebrew University of Jerusalem as the second supervisor of this thesis, and I am gratefully to follow up the research project steps.

I would also like to acknowledge Dr. Siba Shanak, PhD Bioinformatics at the Arab American University-Jenin as the second reader of this thesis and I am gratefully indebted to her for her very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my family for providing me with support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

LIST OF ABBREVIATIONS

ROH	Run Of Homozygosity
IBD	Identical By Descent
GWAS	Whole Genome Association Study
HWE	Hardy Weinberg Equilibrium
MAF	Minor Allele Frequency
SNP	Single Nucleotide Polymorphism
LD	linkage disequilibrium

LIST OF FIGURES

LIST OF TABLES

KEY CONCEPTS:

This thesis includes major concepts related to the genetic terms for ROH in a Palestinian cohort compared with worldwide populations.

Homozygosity: The state of having identical alleles for a specific gene, where these two alleles are inherited from two parents.

Identical-By-Descent (IBD): Identical DNA segments inherited from parents and runs in families.

Inbreeding: A genetic term of consanguineous marriage which increases homozygosity due to IBD and subsequently causes recessive deleterious alleles for the same locus of the gene.

Inbreeding depression: The deleterious effect of inbreeding among the population which reduces variation in the population

Coefficient of inbreeding (F): It is the measure of degree of the two alleles in homozygous state more than being in heterozygous state in individual, because of parental relatedness.

SNP Array: A method for determining polymorphic SNPs build up in the genome.

SNP: Single nucleotide polymorphism- it occurs once for every 300 nucleotides in the human genome. It is the most common variant in the human genome.

GWAS: The study of a sample of population to discover the common genetic variants (SNP) associated with the disease.

Recessive alleles: A weaker allele that is masked by a dominant allele; it can show its phenotype when it is present in both copies (identical alleles).

Recombination: A process, by which pieces of DNA are broken, then rejoined to produce new recombined alleles.

Random segregation: A process of separation of two alleles from each other in the stage of gametes formation; then these alleles reunites in fertilization process.

Meiosis: A process of gametes formation (sperms and eggs). It consists of division of the single cell twice to produce four cells, each having half of the information of the original one.

Haplotype: Set of SNPs inherited together.

Linkage disequilibrium: Tendency for genotypes at two loci to be inherited together.

LOD score: Statistical estimation of location between two genes on the same chromosome.

Genome annotation: The process of adding information of DNA sequences to the genome database, where this information relates to coding regions called genes.

TABLE OF CONTENTS

CHAPTER ONE

1. INTRODUCTION

The offsprings of related parents have high chance of inherit identical chromosomal segments. By this mechanism, the amount of homozygosity increases, while the amount of heterozygoisty decreases. Therefore, the recessive alleles that are hidden by the dominant alleles will be expressed through inbreeding (Alvarez, Quinteiro and Ceballos, 2011). That occurs when the two recessive alleles are in a homozygous situation. (Knapp, 1993). Inbreeding depression is a term used to descibe the detrimental effects of inbreeding, which my result in recessive deletriouse traits (Waller and Keller, 2002). Inbreeding depression occurs in many species of animals and plants as well. In humans, it is caused by the increased homozygosity of individuals. Over-dominance hypothesis or heterozygote advantage also play a minor role in fitness, where the heterozyous allels fitter than homozygous allels. (Charlesworth and Charlesworth, 1999). Partial dominance hypothesis, where the recessive deleterious alleles remain in the population because of mutation (Charlesworth and Willis, 2009). Evidences are available on the effect of paternal relatedness in increasing the risk of complex diseases such as blood pressure and LDL cholesterol and monogenetic disorders (McQuillan et al., 2008). There are many steps that are followed to investigate the effect of inbreeding on the health of an offspring. Therefore, this phenomenon has to be measured at the individual level. Inbreeding coefficient F is the first method used for quantification of this phenomenon (Wright, 1922). This estimation used Wright's path method (Hartl and Clark, 2007). It is based on a calculation of the probability of inheritance of two IBD alleles for each individual. The parent's alleles are transmitted to a specific offspring at 0.5 probability. At that time, there was not an

alternative method for Wright's path method despite its drawbacks (Carothers et al., 2006). In 1980s, scientists used the homozygosity mapping to define many autosomal -recessive genes that cause monogenic human diseases; this explains why the regions surrounding the genes of the disease are IBD in individuals with parental relatedness (Botstein and Lander, 1987). Between 1995 and 2003, there were about 200 studies, they included use of homozygosity mapping for consanguineous marriage to define rare recessive disease gene (Risch and Botstein, 2003). Homozygosity mapping method requires an estimation of the autozygous genome proportion of the affected individual based on LOD score for linkage at a specific locus (Miano et al., 2000), (Leutenegger et al., 2006). Broman and Weber(1999) proposed a method for estimation the genomic data of autozygosity for runs of consecutive homozygous markers. The concept of this method explains that the offspring of the first cousin marriage has identical chromosomal segments from both parents. The identical segments of the first cousin marriages are called IBD or ROH. ROH can be observed using high density genome scan data, which is a more accurate and reliable method for estimation of autozygosity in different individual and in the population (Broman and Weber, 1999). ROH is a new approach developed to identify susceptible loci across the genome. This algorithm is designed to find runs of consecutive homozygous SNPs and their location among a number of samples (Golden Helix, 2016). The study of ROH is significant for both population and medical genetics, where the patterns of ROH are affected by a number of factors such as population bottleneck, population size, and other evolutionary forces such as selective sweeps (Sabeti et al., 2007). Botstein and Lander (1987) mentioned in their study about ROH that the recent parental relatedness prefers to form long ROH about several

mega-bases due to IBD. IBD regions may contain a recessive deleterious allele that would probably cause a genetic abnormality.

1.1. WHAT ARE ROH?

ROH are contiguous segments of homozygous alleles that are present in the human genome because of identical alleles that are transmitted from parents to their offspring (Berry et al., 2012). Also ROH can be defined as large stretches of diploid genome at SNP positions (Nothnagel et al., 2010). Long stretches of consecutive homozygous result from consanguineous marriages, the small size of the population or natural selection (Szpiech et al., 2013). ROH can be statically measured up to 1.5 Mb or more (McQuillan et al., 2008). Magi et al., (2014) define ROH, as being about chromosomal segments that appear in case of homozygosity for diploid genome, where the identical alleles are displayed in the close loci.

1.2. REASONS OF ROH:

ROH occurs due to the transmission of identical alleles (haplotypes) from both parents to their offspring, where the long ROH is due to the recent consanguinity; this means that a recent inbreeding has taken place within a pedigree (McQuillan et al., 2008). Parents that have a recent common ancestor will share a large part of their genome with their offspring that are IBD (Kefi et al., 2015). Kirin et al., (2010) mentioned that the long ROH are due to recent inbreeding because of little rate for recombination that breaks up IBD segmental chromosomes, while ROH, that result from distantly related parents, tend to be shorter because of an increasing rate of recombination that breaks up the chromosomal segments by repeated meiosis.

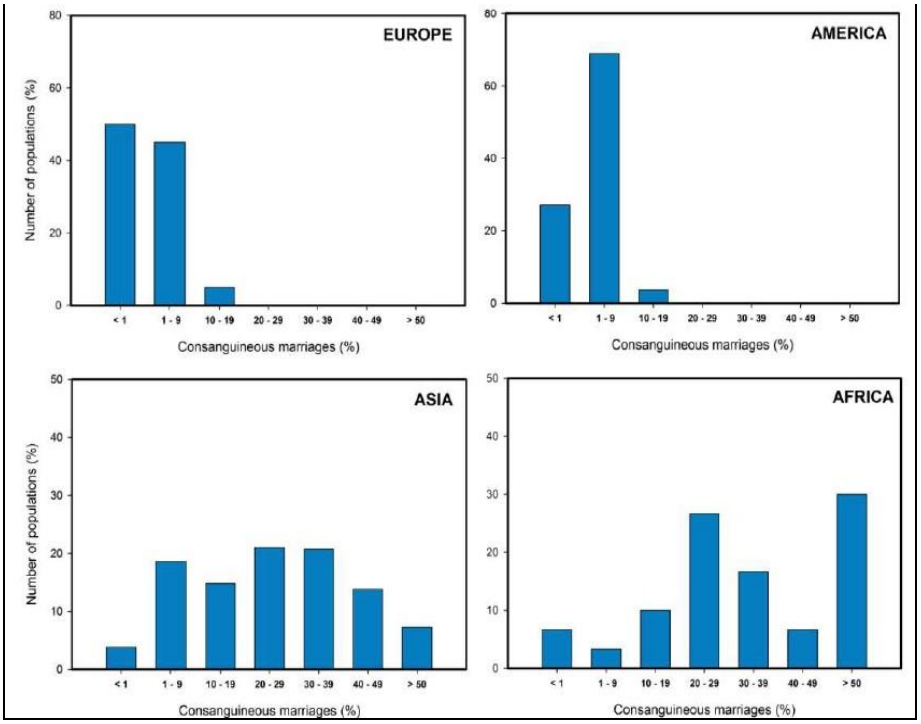
1.2.1. EFFECT OF RECOMBINATION ON IDENTITY BY DESCENT

Previous studies prove that the occurrences of different patterns of ROH for different mammals are linked with recombination rate (Pollinger et al., 2010), (Howrigan, Keller and Simonson, 2011). In non-consanguineous marriages, the IBD segments are broken up by time through recombination processes. For human, ROH disappeared with each generation following an inverse exponential distribution, with common ancestor halving the ROH size (Keller, Visscher and Goddard, 2012), (Nothnagel et al., 2010). That is why size and position of ROH in the genome are likely linked with the rate of recombination. The human genome is enriched with strong LD regions, which are broken up by recombination; which is one of the factors that are influenced by, the length and location of the homozygosity regions as well as its abundance in the genome (Jeffreys, Kauppi and Neumann, 2001). An individual is homozygous for a di-allelic marker if both alleles at that marker are identical. In some individuals, long tracts can be found where homozygous markers occur in an uninterrupted sequence. The most obvious explanation for such tracts is autozygosity, where the same chromosomal segment has been passed to a child from parents who inherited it from a common ancestor. As recombination interrupts long chromosome segments over time, the length of a homozygous segment depends, in part, on the time since the last common ancestor of the parents. Therefore, in an inbred population we expect to see longer homozygous segments than in outbred populations. Scientists have found that homozygous tracts are significantly more common in regions with high linkage disequilibrium and low recombination while the location of tracts is similar across all populations (Gibson, Morton and Collins, 2006).

1.2.2. CONSANGUINITY

ROH are usually investigated in the context of consanguineous marriage, which is still one of the followed and observed traditions in many parts of the world with varying degrees as in figure 1.2.2.1 (Alvarez, Quinteiro and Ceballos, 2011). The figure displays the percentage of consanguineous marriages in four continents. Genetically, consanguineous marriages can be defined as a union between two individuals such as the first and second cousin where F=0.0156 (Hamamy, 2012).

Figure 1.2.2.1: Percentage of consanguineous marriages in human population in four continents



Bennett et al., (2002) defines consanguinity as a term used to define the relationship between two couples who they share at least one common ancestor. The high rate of monogenetic disease comes as result of consanguineous marriage; it causes an increase in the emergence of homozygous recessive alleles; this is more deleterious than other common recessive alleles (Bittles, 2003), (Wright and Andolfatto, 2008). Recently, there is more increased interest in the impacts of inbreeding than in complex genetic disease, such as

cancer, cardiovascular disease and adult diabetes, which lead to mortality and morbidity (Charlesworth and Hughest, 1996).

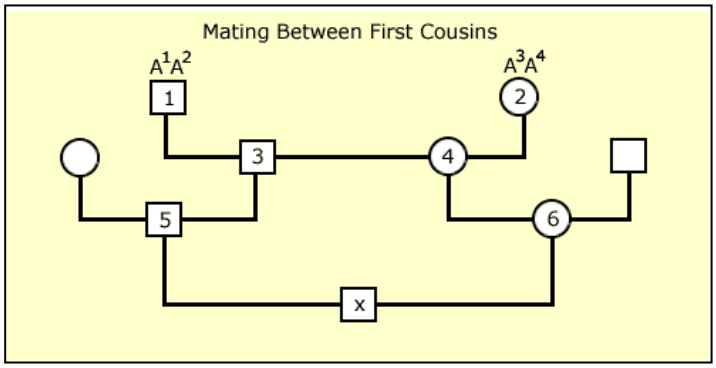
1.2.2.1. CONSANGUINITY IN PALESTINE

In 1995, the percentage of consanguineous marriage was 50% in Palestine, 30% were first cousin marriages. In 2004, this percentage decreased to 45% with 28% was first cousin marriages. This change indicates that the percentage of first cousin marriages is the same over the years, while the percentage of marriages of distant relatives has decreased (Assaf et al., 2009).

1.3. MEAURSING INBREEDING COEFFICIENTS

Several methods are available to calculate the degree of consanguinity between individuals. In general the coefficient inbreeding calculated based on the number of nodes in the pedigree. $F=\frac{1}{2}^{n}$, where n: is the number of nodes in the pedigree. For example to calculate coefficient inbreeding for the offspring X in the figure below.

Figure 1.3.1: Mating between first cousin marriage



Probability that offspring X will receive A^1 through individual 6 = $\frac{1^3}{2} = \frac{1}{8}$

Probability that offspring X will receive A^1 through individual 5 = $\frac{1^3}{2} = \frac{1}{8}$

Probability that offspring X is homozygous for $A^1 = \frac{1}{2}^3 = \frac{1}{8} \times \frac{1}{8} = \frac{1}{64} = 0.015$

In (table 1.3.1), the analysis of coefficient inbreeding using PLINK tool is based on the observed versus expected number of homozygous genotypes; this analysis skips haploid markers (male X and Y chromosome markers),when the value of F approach 1, means high inbreeding (high related parents), while when the value approach 0, means less relatedness between two parents (Medicine, 2005), (Purcell, 2010).

Table 1.3.1: Inbreeding coefficients of Palestinian individuals

FID	IID	F
AK	AK	0.1406
AY	AY	0.158
AZ	AZ	0.1625
BIBJ	BIBJ	0.1552
BX	BX	0.1171
DME	DME	0.08979
DMF	DMF	0.119
EC	EC	0.1912
FHFL	FHFL	0.2913
HYD	HYD	0.1993
IT	IT	0.1335
AJ	AJ	0.1286
BG	BG	0.09409
BK	BK	-0.0516
BQ	BQ	0.1324
C1	C2	0.09762
CJ	CJ	0.03488
CX	CX	0.01449
CZ	CZ	-0.02001
DC	DC	0.08842
DR	DR	0.2054
EA	EA	0.08471
EG	EG	0.08131
EP	EP	0.05804
EQ	EQ	0.02808
GN	GN	0.1037
IB	IB	0.1121
J	J	0.03672
Z	Z	0.02942

1.4. SIZE CLASSIFICATION OF ROH:

The ROH length for each population is modeled into three classes: (A) Short ROH, measuring 10 kb; it results in homozygosity of ancient haplotypes and contributes to local LD patterns. (B) Mediate ROH, measuring 100 kb to one Mb; it results from background relatedness. (C) Long ROH, measuring double of Mb, and it probably results from recent parental relatedness (Pemberton et al., 2012). These runs have harmful variations, where it allows these recessive variants to be expressed and have their effect when presented as both copies in the human genome (Szpiech et al., 2013).

1.5. ROH CONSEQUENCES:

As a result of consanguineous marriages, ROH has a health impact. Inbreeding depression is a term used to express the effect of the consanguineous marriages on health; it refers to a decrease in the fitness of the offspring as a result of parental relatedness (Charlesworth and Willis, 2009). In a study about the association between ROH and Intellectual Disability (ID), Gamsiz et al., (2013) proved that there is a strong association between autosomal homozygsoity and the level of intellectual and adaptive function in the affected probands (Gamsiz et al., 2013). Moreover, in the autism involving ID, there is an increase in the ROH of the probands compared with unaffected siblings (Kuningas et al., 2011). This study finding was confirmed by another study which aimed to reveal important roles of autosomal recessive variants in intellectual disability; the amounts of homozygsoity modify the degree of cognitive impairment despite the causes of intellectual disability (Ilaria et al., 2014). In addition, sum of lengths of ROH is linked with decreasing the stature; is reflected in the measurement of the lung functions. Additionally, it causes a reduction of the

cognitive function when the educational level and other cognitive factors are measured (Joshi, et al., 2014). Concerning the relationship between ROH and coronary diseases, it has been suggested that the accumulation of the recessive variants may increase the risk of coronary artery disease (Christofidou, et al., 2015). As for the association between ROH and Alzheimer disease, an increase was detected in the amount of ROH in outbreed patients of Alzheimer disease in African and American populations (Mahdi, et al., 2015). However no associations with individual ROH were with schizophrenia (Lencz et al., 2007), bipolar (Vine et al., 2009) colorectal cancer (Spain et al., 2009) and childhood acute lymphoblastic leukemia (Hosking et al., 2010).

CHAPTER TWO

2. PROBLEM STATEMENT

Consanguineous marriages are one of the common phenomena that provides risk for genetic diseases. Palestine is considered one of the developing countries that noticeably have this phenomenon, mainly in the southern part of West Bank. The prevalence of monogenetic disease, as a result of inbreeding, has long been recognized. Consequently, there is an obvious lack of health awareness of the issue of consanguineous marriages.

2.1.AIMS AND OBJECTIVES OF THE THESIS:

In the context of consanguineous marriage, ROH are usually investigated. In small isolated populations, immigration rates are so low, so cousin marriage is inevitable. The degree of kinship between two individuals in a population depends directly on the size of this population; thus, mating between pairs of individuals is inevitable, and it is more closely related for smaller population (Hill and Mackay, 2004).

This study has three objectives:

- To define ROH in a Palestinian cohort.
- To see whether these runs segments are shared by other populations.
- To gene map these segments.

CHAPTER THREE

3. MATERIALS AND METHODS:

The study methodology aimed to extract ROH segments from Palestinian samples. The participants gave their prior approval to be part of this study. Homozygosity mapping of DNA taking from all participants. This step utilized SNP array (250K). Computational methods were conducted to reach the final outcomes. The ethical aspects have been observed throughout the entire study.

3.1. ETHICAL CONSIDERARTION

Bethlehem University approved the study following a prior consent of all families involved.

3.2. PARTICIPANTS

Palestinian samples taking part in this study consisted of 29 individuals from 29 families. Samples were taken from the population of one country. The original sample contained 43 individuals. This sample did not represent the entire Palestinian population.

3.3. HOMOZYGOSITY MAPPING:

According to the protocol of the Affymetrix GeneChip(http://www.affymetrix.com), the DNA samples for individuals of this study was hybridized on 250K SNP chip array. The array were scanned using the Affymetrix Gene Chip Scanner 3000 with the Gene Chip Operating System 1.4 (GCOS). Data were analyzed using the Gene Chip DNA Analysis Software 2.0 (GDAS) and classified by chromosome and sorted on the basis of

the physical position of the SNPs. Homozygous regions were defined as segments where SNPs were homozygous for one allele among affected relatives but discordant (heterozygous or homozygous of the complement allele) among unaffected relatives in the pedigree. Enhanced stringency requirements for SNP quality were applied, with only SNPs with a quality score better than 0.06 from GDAS and GCOS being retained, thus using about 67% of the SNPs in homozygosity analyses. A segment was considered continuously homozygous if there was no more than one heterozygous SNP in any 10-SNP window. Homozygous segments of 2 Mb or longer were retained for further analysis. LOD scores12 were calculated for these homozygous segments for each family, exploiting genotypes of all informative SNPs. The LOD score for a family depends on the family's structure and parameters of the recessive model and is the same for all informative homozygous regions in the genome **(Shahin et al., 2010)**.

3.4. PLINK TOOLS FOR ROH ANALYSIS:

PLINK is an open source software used as GWA toolset for genetic analysis procedures; it is used here for ROH analysis. For this procedure, data should be prepared in two files; one for the genetic information and the other one for the individuals to be analyzed (Purcell, 2010). In this study, data files for individuals were merged into these two types. To this aim, Perl languages were used. 'MAP file', which is space delimited file, contains four columns. The file has information about the location of the SNP markers used in the analysis; the detailed format for which is found in (Table 3.4.1):

Table 3.4.1: Example of MAP file

Chromosome	SNP ID	Genetic Dist.	Physical position
1	rs123456	0	1234567
1	rs223456	0	1234568
1	rs243456	0	1234569

Each row represents information for each SNP in the genome. The genetic distance column was not used. Physical position (column 4) is the location of SNP in the chromosome.

The other file (PED file), is also space -or tab-delimited, which contains six columns (Purcell, 2010). This file has the following information of genotype data, as the following format, in the (table 3.4.2) below:

Table 3.4.2: Example of PED file

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Phenotype	Genotype
12345	12345	0	0	1	2	A
22345	33245	0	0	2	1	A
32345	32345	0	0	1	2	A

Each row represents data for one individual. Family ID and Individual ID are the same. Parental ID is unknown and was not used in this analysis (both paternal and maternal, column 3 and 4, respectively); column 5 identifies the gender of the individual. Phenotype is used for the disease status (1=unaffected, 2=affected). The last column represents the bi-allelic set of basis with all the individuals SNP markers (Purcell, 2010).

3.4.1. ROH ANALYSIS:

There are five important steps in this analysis, three of which uses PLINK:

1. **Remove very closely related individuals**: The aim of this analysis is to look for extended haplotypes shared between distant relatives because having individuals of close relatives, such as siblings or first cousins is likely to swamp the results of the analysis. This can be done through using the following command to generate the genome file in (table 3.4.1.1) below:

```
Plink --file mydata --genome
```

Table 3.4.1.1: Genome

FID1	Family ID for first individual
IID1	Individual ID for first individual
FID2	Family ID for second individual
IID2	Individual ID for second individual
RT	Relationship type given PED file
EZ	Expected IBD sharing given PED file
Z0	P(IBD=0)
Z1	P(IBD=1)
Z2	P(IBD=2)
PI_HAT	P (IBD=2)+0.5*P(IBD=1) (proportion IBD)
PHE	Pairwise phenotypic code (1,0,-1 = AA, AU and UU pairs)
DST	IBS distance (IBS2 + 0.5*IBS1) / (N SNP pairs)
PPC	IBS binomial test
RATIO	Of HETHET: IBS 0 SNPs (expected value is 2)

Through scanning of these files, a- check is performed on whether any individual has high PI_HAT (more than 0.05), where PH_HAT is calculated based on IBD. To reduce the file size, the –min X option is used to only output to plink genome pairs where PI HAT is greater than X.

Thus, the command will be:

Plink --file mydata --genome --min 0.05

2. **Prune the set of SNPs**: This analysis requires approximately independent SNPs in order to use this command for pruning :

```
plink --bfile mydata2 --mind 0.1 --geno 0.1 –hwe 0.01 --make-bed --out mydata3
```

--mind: This command is used to remove individuals with high missing genotype.

--geno: This command is used to filter out the missing call rate which is more than the default value (0.1).

--HWE: This command is used to exclude markers that fail hardyweinberg test at specific threshold.

3. **Define ROH segments**: Using PLINK commands, which depend on simple screen for ROH within an individual, we define ROH for certain number of SNPs spanning for certain KB distance. The algorithm in PLINK starts by taking a window of X SNPs and sliding it across the genome. Each window position will allow some heterozygous or missing calls, and it can be determined if they look homozygous. For each SNP algorithm, we can calculate the rate of homozygosity window that overlaps with that position. The exact window size and the threshold that is related to SNP density, as well as with the expected size for the ROH are important to dense SNP maps and scan large segments. In this algorithm, the long ROH are not broken up using heterozygote. This approach can be also used for population parameters such as allele frequency and recombination rate. The commands used for defining ROH:

   plink --file pal --homozyg --homozyg-kb 1000 --homozyg-snp 10 --homozyg-density 100 --homozyg-gap 10000 --homozyg-group --homozyg-match 0.9 --homozyg-

window-threshold 0.1 --homozyg-window-missing 10 --homozyg-window-het 10 --out tmp1.

- **--homozyg-window-snp:** This command is used to specify window size.
- **--homozyg-window-snp 50:** This command is used to specify that the window is exactly 50 SNPs long.
- **--homozyg-window-het:** This command is used to specify the number of heterozygote's allowed for each window.
- **--homozyg-window-missing:** This command is used to set the number of missing calls allowed within a window.
- **--homozyg-window-threshold:** This command is used to specify the proportion of overlapping windows that must be called as homozygous for any given SNP within those windows to be considered homozygous.
- **--homozyg-density:** This command is used to specify the density of SNPs required for some range of KB.
- **--homozgy-gap:** This command is used to split the segment into two pieces.
- **--homozyg-window-snp and --homozyg-window-kb:** These are the final commands used to determine whether the segment is considered part of ROH.

In ROH analysis, the PLINK moves sliding windows for specific length for each individual genotype for each SNP marker. Also, the window moves from one position to another. The % of homozygosity is recoded for each SNP position. This is the window size hypothesis for the 5 SNPs in a row, which are represented visually as in (table 3.4.1.2) below:

Table 3.4.1.2: Sliding window algorithm

position12												4/5	4/5	4/5	4/5	4/5
position11											5/5	5/5	5/5	5/5	5/5	
position10										5/5	5/5	5/5	5/5	5/5		
position9									4/5	4/5	4/5	4/5	4/5			
position8								4/5	4/5	4/5	4/5	4/5				
position7							4/5	4/5	4/5	4/5	4/5					
position6						3/5	3/5	3/5	3/5	3/5						
position5					2/5	2/5	2/5	2/5	2/5							
position4				3/5	3/5	3/5	3/5	3/5								
position3			3/5	3/5	3/5	3/5	3/5									
position2		3/5	3/5	3/5	3/5	3/5										
position1	4/5	4/5	4/5	4/5	4/5											
zygosity >	**hom**	**hom**	**hom**	**hom**	**het**	**het**	**hom**	**hom**	**het**	**hom**	**hom**	**hom**	**hom**	**hom**	**hom**	**het**
SNP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**

PLINK allows the sliding window to be defined as a specified number of SNPs in a row. The output file would contain 13 columns, and we need to extract consensus rows with a specific condition; the tenth column is about the size of ROH. For our results the size should be >> 1000kb, and it should be shared by at least 15 out of 29 individuals. Accordingly, the following result would be obtained. The table describes the segments that are meeting the conditions, where ROH size >1000KB and the number of individuals >15.

3.5. BEDTOOLS FOR GENOMIC ARITHMETIC

BEDTOOLS is a collection of tools used for a wide range of genomic analysis, such as intersection, subtraction, merge, count, and complementation. The file format must be as BAM, BED, GFF/GTF, and VCF. The analysis was done for these operations using LINUX command line (Quinlan and Kindlon, 2016). Table 3.5.1, represents the number of individuals for each population that has shared ROH segments with Palestinian population in the three chromosomes. In this study, it is needed to check for any intersection between each

segment for the Palestinian population with the corresponding segments on the same chromosome along the 53 -CEPH populations. Intersection algorithm works as follows:

Table 3.5.1: BED file for Palestinian

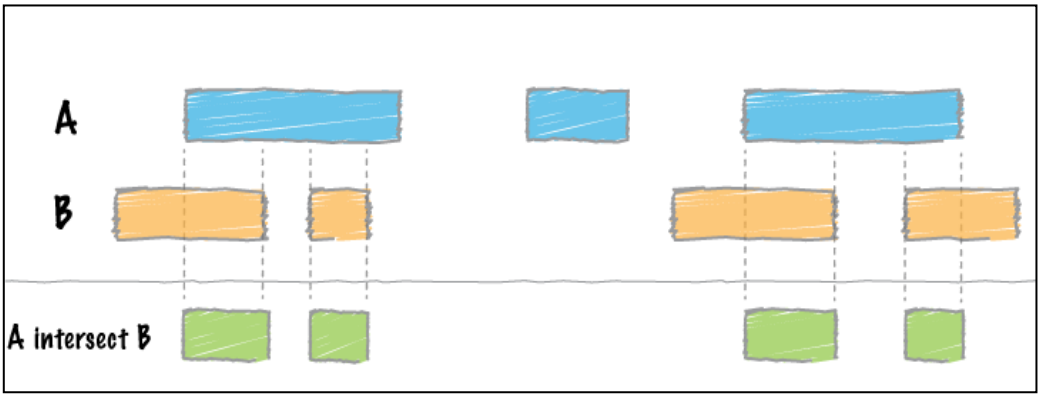
individuals	Chromosome	Start	End
23	Chromosome 15	72224840	73499534
23	Chromosome 8	6909287	8238782
18	Chromosome 5	279376	1322006
16	Chromosome 5	68355443	70739833
16	Chromosome 5	45752153	49631829

In general, the BED file contains the chromosome name, start position of the SNP and the end position of the SNP on that chromosome. Using the default behavior for the BEDTOOLS intersection, overlapping of segments will be detected; this tool reports the shared intervals between the two overlapping features, as in the following example:

```
$ cat All.bed
chr1  10  20
chr1  30  40
$ cat Pal.bed
Chr1  15  20
$ bedtools intersect –a Pal.bed –b All.bed
Chr1 15 20
```

This example works as follows (figure 3.5.1):

Figure 3.5.1: Bedtool intersect function



3.6. DATABASE

For this study, a database has been created for the all populations, bed files that were taken from Pemberton Lab (Pemberton, 2015). This database contains a table for 53 populations. The tables contain the information of bed file, start SNP, end SNP, and the chromosome number.

OMIM GENE MAP SEARCH: This step aims to search for genomic gene regions on the chromosome to receive all the genes of OMIM and the loci of those regions. Using OMIM search map, it was found that the long arm of chromosome 15 has many genes linked with diseases as in the table 3.7.1. The Hexosaminidase A alpha polypeptide is a gene that encodes a member of glycosyl hydrolase 20 families of proteins. Mutation of this gene leads to neurodegenerative disorders such as Tay- Sachs disease (NCBI, 2016).

Gene Map Search – 'chr15:72224840-73499534'

Table 3.7.1: 'chr15:72224840-73499534'

Genomic coordinates	Gene/	Gene/	Comments	Phenotype	Inheritance
15:19000000-101991189	HCVS	Human coronavirus			
15:19000000-101991189	HYT2	Hypertension,		{Hypertension,	Multifactorial
15:19000000-101991189	LCS1, CHLS	Cholestasis-		Cholestasis-	Autosomal
15:58800000-78000000	GLC1N	Glaucoma 1, open	max lod at	Glaucoma 1,	
15:58800000-72400000	USH1H	Usher syndrome, type	max lod at	Usher	
15:67200000-101991189	GLM4	Glioma susceptibility 4	max lod at	{Glioma	
15:72199028-72231623	PKM, PKM2,	Pyruvate kinase,			
15:72284726-72320183	CELF6,	CUGbp- and ELAV-			
15:72343434-72376472	HEXA, TSD	Hexosaminidase A,	on 15q+ in APL	GM2-	Autosomal
15:72343434-72376472	HEXA, TSD	Hexosaminidase A,	on 15q+ in APL	[Hex A	Autosomal
15:72343434-72376472	HEXA, TSD	Hexosaminidase A,	on 15q+ in APL	Tay-Sachs	Autosomal
15:72400000-88500000	CILD8	Ciliary dyskinesia,	max lod at	Ciliary	
15:72400000-78000000	DEL15q24,	Chromosome 15q24	contiguous gene	Chromosome	Isolated cases
15:72400000-78000000	ENFL2	Epilepsy, nocturnal	some ENFL not	Epilepsy,	Autosomal
15:72400000-78000000	MRST	Mental retardation,		Mental	
15:72474281-72586562	ARIH1, ARI,	Ariadne, Drosophila,			
15:72686178-72738475	BBS4	BBS4 gene		Bardet-Biedl	Autosomal
15:72751366-72783784	ADPGK	ADP-dependent			
15:73051714-73305205	NEO1, NGN	Neogenin, chicken,			
15:73319858-73369263	HCN4, SSS2	Hyperpolarization-		Sick sinus	Autosomal
15:73319858-73369263	HCN4, SSS2	Hyperpolarization-		Brugada	
	COL1AR	Collagen I, alpha,			

Gene Map Search - 'chr8:6909287-8238782'

Table 3.7.2:chr8:6909287-8238782

Genomic coordinates	Gene	Gene/Locus name	Comments	Phenotype	Inheritance
	DBA2	Diamond-Blackfan		Diamond-	
	GEFSP6	Generalized epilepsy	between	Epilepsy,	
	KWE	Keratolytic winter		Keratolytic	Autosomal
	MYP10	Myopia 10		Myopia 10	Multifactorial
	TDH	L-threonine			
	WS2C	Waardenburg		Waardenburg	
8:6300000-12800000	DIH2	Hernia, congenital		Hernia,	Autosomal
8:6300000-12800000	SLEB12	Systemic lupus	associated with	{Systemic lupus	
8:6924693-6926075	DEFA6, DEF6	Defensin, alpha-6,			
8:6935819-6938337	DEFA4,	Defensin, alpha-4,			
8:6977648-6980118	DEFA1,	Defensin, alpha-1,			
8:7055299-7056738	DEFA5, DEF5	Defensin, alpha-5,			

8:7255242-7260475	FAM90A15	Family with sequence	copy 1		
8:7262865-7268097	FAM90A3	Family with sequence	copy 2		
8:7278109-7283341	FAM90A13	Family with sequence	copy 4		
8:7285731-7290963	FAM90A5	FAmily with sequence	copy 5		
8:7293353-7298583	FAM90A20	Family with sequence	copy 6		
8:7428887-7430347	DEFB103A,	Defensin, beta 103A			
8:7447753-7463669	SPAG11,	Sperm-associated			
8:7556137-7559103	FAM90A7	Family with sequence	copy 8		
8:7713782-7719014	FAM90A14	Family with sequence	copy 12		
8:7721428-7726662	FAM90A18	Family with sequence	copy 13		
8:7736724-7741957	FAM90A8	Family with sequence	copy 15		
8:7752019-7757253	FAM90A19	Family with sequence	copy 17		
8:7759667-7764901	FAM90A9	Family with sequence	copy 18		
8:7769583-7771312	FAM90A10	Family with sequence	copy 19		
8:7894564-7896715	DEFB4A,	Defensin, beta-4a			
8:8027073-8032304	FAM90A12	Family with sequence	copy 22		
	AIS3,	Autoimmune disease,		{Autoimmune	
	IFNB3	Interferon, beta-3,	previously		
	ZNF1	Zinc finger protein-1			

Gene Ma p Search - chr5:'279376-1322006'

Table 3.7.3:chr5:279376-1322006

Genomic coordinates	Gene	Gene/Locus name	Comments	Phenotype	Inheritance
	ASD1	Atrial septal defect 1	max lod at	Atrial septal	Autosomal
	BCC3	Basal cell carcinoma,		{Basal cell	
	GLM8	Glioma susceptibility 8	associated	{Glioma	
	LNCR3	Lung cancer	associated	{Lung cancer	
	LSINCT5	Long stress-induced			
	MCDR3	Macular dystrophy,	maximum lod at	Macular	Autosomal
	MHS6	Malignant		{Malignant	
	MYP16	Myopia 16	max lod at	Myopia 16	
	TST2	Tuberculin skin test	max lod at	[Tuberculin skin	
5:271620-314973	PDCD6,	Programmed cell death			
5:304175-438290	AHRR,	Arylhydrocarbon			
5:443183-471937	EXOC3, SEC6	Exocyst complex			
5:471985-524433	SLC9A3,	Solute carrier family 9	pseudogene on	Diarrhea 8,	Autosomal
5:612289-663499	CEP72,	Centrosomal protein,			
5:659861-730453	TPPP, P25,	Tubulin			
5:892853-919347	TRIP13,	Thyroid hormone			
5:1008844-1038811	NKD2	Naked cuticle,			
5:1050373-1155886	SLC12A7,	Solute carrier family			
5:1201594-1225116	SLC6A19,	Solute carrier family 6		Iminoglycinuria,	Autosomal
5:1201594-1225116	SLC6A19,	Solute carrier family 6		Hyperglycinuria	Autosomal
5:1201594-1225116	SLC6A19,	Solute carrier family 6		Hartnup	Autosomal

5:1225354-1246188	SLC6A18,	Solute carrier family 6			
5:1253166-1295625	TERT, TCS1,	Telomerase reverse	deleted in cri du	{Dyskeratosis	Autosomal
5:1253166-1295625	TERT, TCS1,	Telomerase reverse	deleted in cri du	{Dyskeratosis	Autosomal
5:1253166-1295625	TERT, TCS1,	Telomerase reverse	deleted in cri du	{Pulmonary	Autosomal
5:1253166-1295625	TERT, TCS1,	Telomerase reverse	deleted in cri du	{Melanoma,	
5:1253166-1295625	TERT, TCS1,	Telomerase reverse	deleted in cri du	{Leukemia,	Autosomal
5:1317743-1345069	CLPTM1L,	CLPTM1-like protein			

Gene Ma p Search - 'chr5:68355443-70739833'

Table 3.7.4:chr5:68355443-70739833

Genomic coordinates	Gene	Gene/Locus name	Comments	Phenotype	Inheritance
5:48800000-181538259	BSZQTL2	Bone size quantitative		{Bone size	
5:59600000-93000000	EJM4	Myoclonic epilepsy,	max lod at	Myoclonic	Autosomal
5:67400000-93000000	AAT2, FAA2	Aortic aneurysm,		Aortic	
5:67400000-115900000	GINGF2,	Fibromatosis, gingival,	formerly	Fibromatosis,	
5:67400000-77600000	TSHQTL1	Thyroid-stimulating	associated with	[Thyroid-	
5:69093948-69131071	SLC30A5,	Solute carrier family 30			
5:69167009-69178244	CCNB1	Cyclin B1			
5:69189547-69210356	CENPH	Centromeric protein H			
5:69217745-69230157	MRPS36	Mitochondrial	6 pseudogenes		
5:69234794-69277429	CDK7, STK1,	Cyclin-dependent	previously		
5:69273086-69333068	CCDC125,	Coiled-coil domain-			
5:69364742-69370012	TAF9,	TAF9 RNA			
5:69369296-69414800	RAD17	RAD17, S. pombe,			
5:69415111-69444021	MARVELD2,	Marvel domain-		Deafness,	Autosomal
5:69492291-69558103	OCLN,	Occludin		Band-like	Autosomal
5:70049522-70077594	SMN2	Survival of motor		{Spinal	Autosomal

Gene Map Search - 'chr5:45752153-49631829'

Table 3.7.5:chr5:45752153-49631829

Genomic coordinates	Gene	Gene/Locus name	Comments	Phenotype	Inheritance
5:48800000-181538259	BSZQTL2	Bone size quantitative		{Bone size	
5:59600000-93000000	EJM4	Myoclonic epilepsy,	max lod at	Myoclonic	Autosomal
5:67400000-93000000	AAT2, FAA2	Aortic aneurysm,		Aortic	
5:67400000-115900000	GINGF2,	Fibromatosis, gingival,	formerly	Fibromatosis,	
5:67400000-77600000	TSHQTL1	Thyroid-stimulating	associated with	[Thyroid-	
5:69093948-69131071	SLC30A5,	Solute carrier family 30			
5:69167009-69178244	CCNB1	Cyclin B1			
5:69189547-69210356	CENPH	Centromeric protein H			
5:69217745-69230157	MRPS36	Mitochondrial	6 pseudogenes		

5:69234794-69277429	CDK7, STK1,	Cyclin-dependent	previously		
5:69273086-69333068	CCDC125,	Coiled-coil domain-			
5:69364742-69370012	TAF9,	TAF9 RNA			
5:69369296-69414800	RAD17	RAD17, S. pombe,			
5:69415111-69444021	MARVELD2,	Marvel domain-		Deafness,	Autosomal
5:69492291-69558103	OCLN,	Occludin		Band-like	Autosomal
5:70049522-70077594	SMN2	Survival of motor		{Spinal	Autosomal

3.7. DAVID Functional Annotation Tool

This step aims to detect and identify the locations of all the genes and all the encoding regions in the genome to determine the function of these genes. Once the genome is sequenced, it is needed to make annotation in order to make sense of the sequenced genome.

Table 3.8.1.1: A 2x2 contingency table for P53 signaling pathway genes

	User Genes	Genome
In Pathway	3-1	40
Not In Pathway	297	29662

Table 3.8.1: Functional Annotation chart Report

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	SMART	DEFSN	RT		5	1.2	2.5E-10	2.7E-9
	INTERPRO	Mammalian defensin	RT		5	1.2	7.8E-10	5.6E-8
	SP_PIR_KEYWORDS	defensin	RT		6	1.4	1.1E-8	1.1E-6
	PIR_SUPERFAMILY	PIRSF001875:alpha-defensin	RT		4	0.9	5.4E-8	6.5E-7
	INTERPRO	Alpha defensin	RT		4	0.9	6.4E-8	2.3E-6
	INTERPRO	Defensin propeptide	RT		4	0.9	6.4E-8	2.3E-6
	INTERPRO	Alpha-defensin	RT		4	0.9	6.4E-8	2.3E-6
	SP_PIR_KEYWORDS	antibiotic	RT		6	1.4	1.6E-7	8.3E-6
	SP_PIR_KEYWORDS	Antimicrobial	RT		6	1.4	2.0E-7	6.9E-6
	SP_PIR_KEYWORDS	fungicide	RT		4	0.9	9.8E-7	2.5E-5
	GOTERM_BP_FAT	defense response to fungus	RT		4	0.9	2.8E-6	1.1E-3
	GOTERM_BP_FAT	killing of cells of another organism	RT		4	0.9	3.5E-6	7.2E-4
	GOTERM_BP_FAT	defense response to bacterium	RT		6	1.4	4.3E-6	5.9E-4
	GOTERM_BP_FAT	response to fungus	RT		4	0.9	1.9E-5	2.0E-3
	GOTERM_BP_FAT	cell killing	RT		4	0.9	1.9E-5	2.0E-3
	GOTERM_BP_FAT	response to bacterium	RT		6	1.4	6.0E-5	4.9E-3
	SP_PIR_KEYWORDS	homodimer	RT		4	0.9	6.1E-4	1.2E-2
	PIR_SUPERFAMILY	PIRSF001878:crotamine	RT		2	0.5	5.7E-3	3.4E-2
	INTERPRO	Sodium	RT		2	0.5	9.6E-3	2.1E-1
	GOTERM_BP_FAT	defense response	RT		6	1.4	1.1E-2	5.2E-1

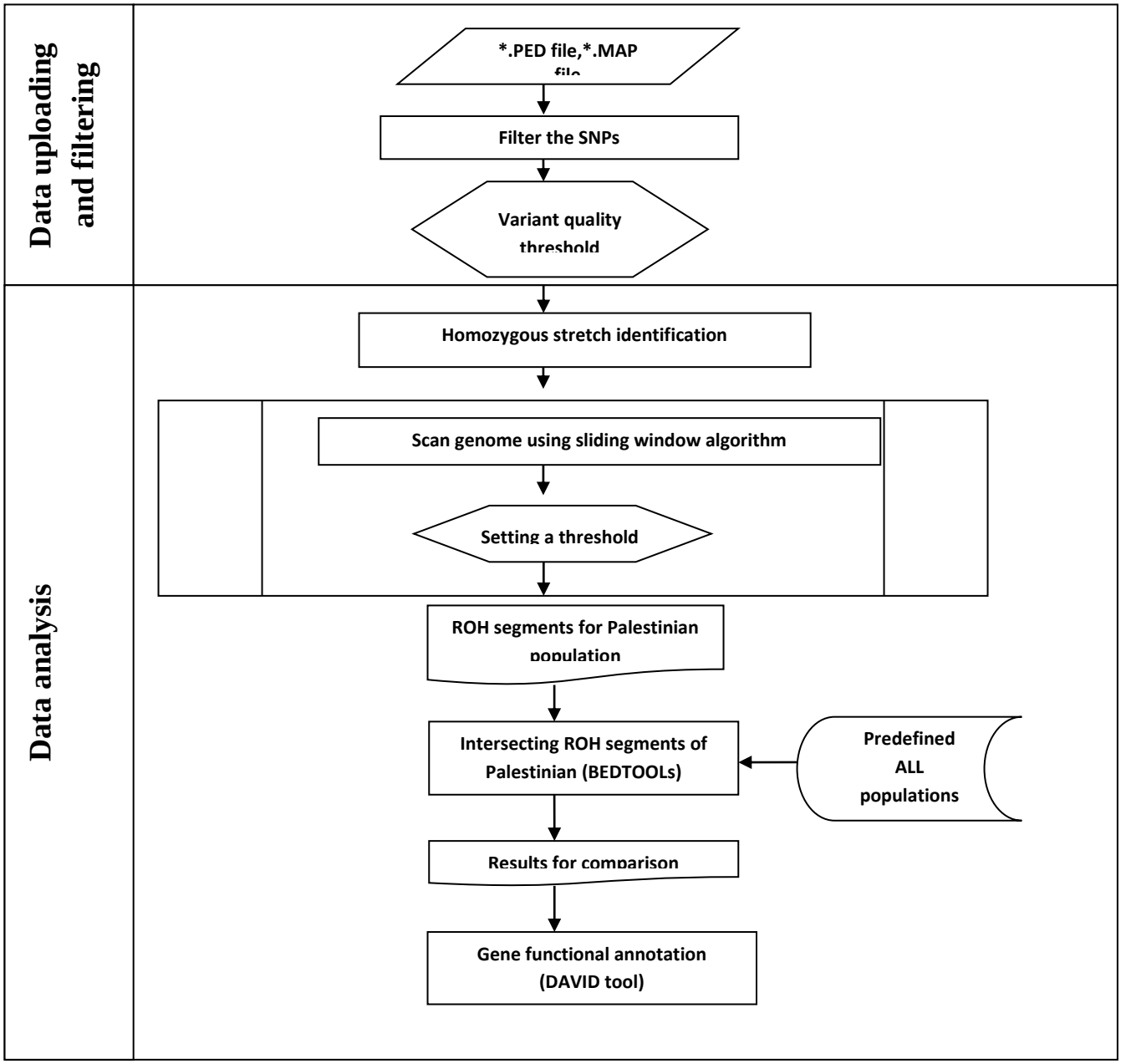
In the above (table 3.8.1), the threshold count represents the minimum number of genes for the corresponding term. P-value represents the maximum score. The category column represents the original database or the resource where the term heads. The term column represents the enriched terms associated with the gene list. *P-value* column is a modified Fisher's exact *p*-value or Ease score, where a smaller value indicates higher enrichment.

3.7.1. **Functional Annotation Chart Report:** This report lists annotation terms associated with the genes. This report only displays significant calculated values (*p*-value), using Fisher's exact statistical test. Fisher test is used to determine gene enrichment in annotation terms. For example, in the human genome, there are 30,000 genes as a total. Forty of them are involved in p53 signaling pathway. Let us suppose that the required list of genes has 300 genes, and 3 of them belong to p53 signaling pathway. We ask if 3/300 is more than 40/30000, *P-value* is equal to 0.008<=0.01; this means that this list of genes is enriched with p53 signaling pathway genes (Huang, Sherman and Lempicki, 2009).

The diagram below (Figure 3.8.1) summaries the computational steps for identifying ROH segments. The process starts with data uploading and filtering steps which consist of uploading .PED and .MAP files, and then filtering the SNPs. This figure is about the flowchart of the followed computerized protocol in identifying ROH in a Palestinian cohort. This step starts with uploading and filtering the data to identifying ROH segments, that was using PLINK software. This software requires two files, PED and MAP file. It needs to filter the SNPs by determining threshold values such as allowing a number of heterozygosis and some missing genotype. In the stage of data analysis, the segment is identified after making

many trials by trying new criteria of windows algorithm in the PLINK to get Palestinian ROH segments. After that we can get bed file of Palestinian cohort to compare it with worldwide population to get the shared segments that was using BEDTOOLs. The final step is to use DAVID tool in order to determine gene enrich segments.

Figure 3.8.1: Identifying ROH workflow



CHAPTER FOUR

4. RESULTS

This study has been conducted to measure the length of ROH for Palestinian individuals, mainly from southern part, based on the fact that they have parental relatedness among their population. The individuals have been selected and filtered using bioinformatics tools where there is no relatedness between them.

4.1. REMOVE VERY CLOSE INDIVIDUALS

Finding ROH segment requires the elimination of very close individuals. We started our project with 43 individuals with 202152 SNPs. After using the plink commands, we ended up with 29 individuals with 197701 SNPs.

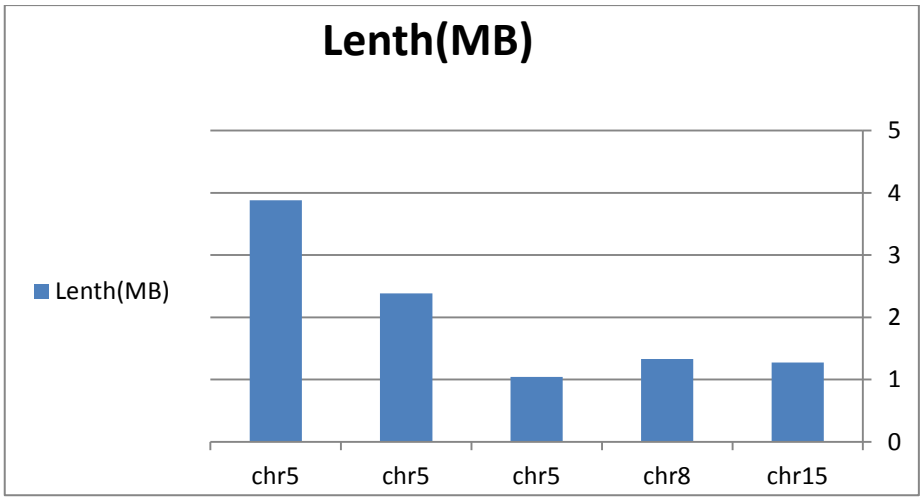
4.2. DEFINE ROH SEGMENTS

After the plink commands are used to define ROH segments, the following results were found as in table 4.2.1. The length of ROH>=1MB and the number of individuals shared these segments= 15 individuals. Table(4.2.1) represents ROH segments in the Palestinian individuals taking part in this study. Figure 4.2.1: ROH segments for 15 out of 29 individuals; the length of ROH is greater than 1MB. The ROH segments were found in chromosome 5 in three regions, with 1.04, 2.3, 3.8 MB length respectively. Nevertheless, in chromosome 8 there is one segment with 1.3 MB in length. In chromosome 15, there is one segment with 1.2 MB in length.

Table 4.2.1: Palestinian ROH Segments

Segment	individuals	Chr	SNP1_ID	SNP2_ID	Start	End	KB	#SNP
S31	23	15	SNP_A-2031697	SNP_A-1947703	72224840	73499534	1274.69	24
S43	23	8	SNP_A-4235252	SNP_A-1886104	6909287	8238782	1329.49	17
S1112	18	5	SNP_A-2044242	SNP_A-2081590	279376	1322006	1042.63	22
S2869	16	5	SNP_A-4202432	SNP_A-1853032	68355443	70739833	2384.39	17
S2873	16	5	SNP_A-1981477	SNP_A-1828858	45752153	49631829	3879.68	13

Figure 4.2.1: Palestinian ROH Segments



4.3. COMPARISON OF ROH SEGMENTS

The aim of this step is to hold a comparison between ROH segments of our Palestinian population sample with that of worldwide populations to get shared segments that are common between them. Table 4.3.1 below displays the numbers of shared segments after making a comparison between ROH segments in worldwide populations and those found in the Palestinian population. The worldwide population bed files were extracted from Pemberton lab database (Pemberton, 2015). In this study, Fifty three out of sixty four were used for comparison.

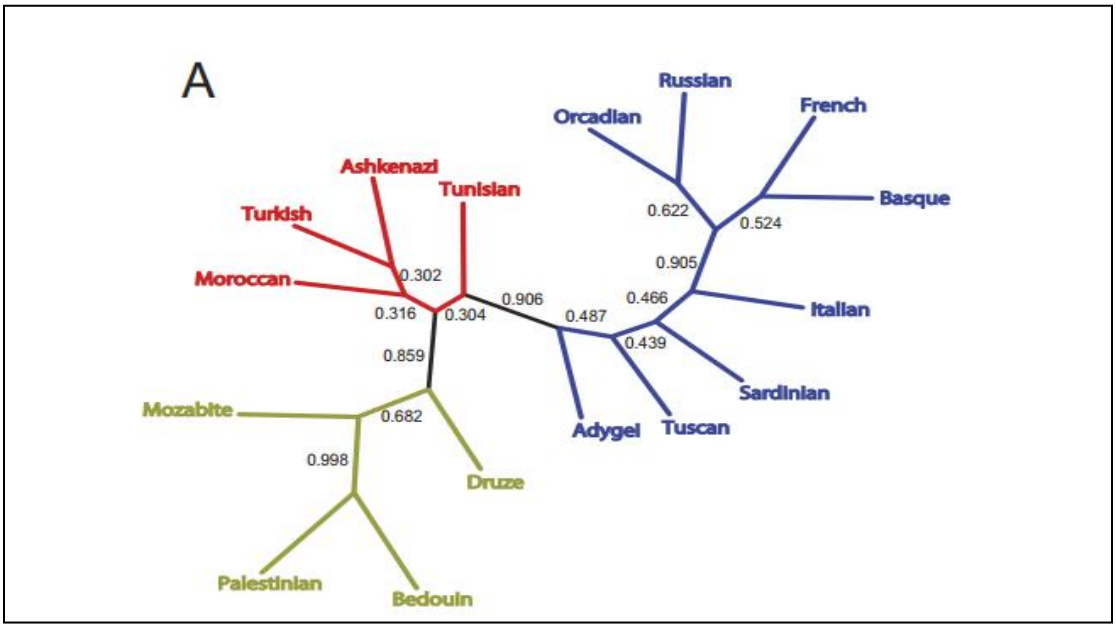
Table 4.3.1: Comparison between Palestinian population and all worldwide

Population	Chr15	Chr5-1	Chr5-2	Chr5-3	Chr8
Palestinian	23	18	16	16	23
Adygei	7	-	2	6	-
Balochi	-	5	4	7	-
Bantu_N.E.	-	1	-	-	-
Bantu_S.E._Tswana	-	-	-	1	-
Bantu_S.W._Herero	-	1	-	-	-
Biaka_Pygmy	-	1	2	6	2
Brahui	-	4	5	8	4
Burusho	-	3	4	7	2
Cambodian	-	1	1	2	-
Colombian	-	2	3	4	3
Dai	-	1	2	5	2
Daur	-	3	2	5	3
French	-	3	8	10	4
French_Basque	-	5	6	10	6
Han	-	9	9	16	8
Hazara	-	2	5	11	4
Hezhen	-	3	2	4	1
Japanese	-	7	10	7	11
Kalash	-	3	7	12	5
Karitiana	-	4	10	3	6
Lahu	-	-	-	-	2
Makrani	-	4	8	8	3
Mandenka	-	1	-	2	1
Maya	-	5	7	10	3
Mbuti_Pygmy	-	-	-	2	1
Miaozu	-	3	2	6	2
Mongola	-	-	3	3	2
NAN_Melanesian	-	2	-	6	6
Naxi	-	2	3	5	2
North_Italian	-	2	2	3	2
Orcadian	-	3	2	8	4
Oroqen	-	1	1	1	2
Papuan	-	10	9	5	11
Pathan	-	1	2	7	7
Pima	-	3	2	4	8
Russian	-	1	4	11	3
San	-	1	1	2	-
Sardinian	-	2	4	12	2
She	-	3	1	3	1
Sindhi	-	1	5	6	-

Surui	**-**	**2**	**4**	**3**	**2**
Tu	**-**	**2**	**1**	**4**	**4**
Tujia	**-**	**1**	**2**	**4**	**3**
Tuscan	**-**	**1**	**2**	**1**	**-**
Uygur	**-**	**2**	**1**	**5**	**1**
Xibo	**-**	**-**	**-**	**5**	**3**
Yakut	**-**	**4**	**3**	**15**	**-**
Yizu	**-**	**2**	**1**	**5**	**4**
Yoruba	**-**	**-**	**2**	**1**	**-**

Significantly, the Adyghe is the only population that has seven segments in chromosmome15 that are shared with the Palestinian population. Figure 4.3.1 represents the relationship between the three types of populations, Middle Eastern (Moroccan, Turkish, Ashkenazi, Tunisian), Jewish (Mozabite, Bedouin, Druze) and European (Adyghe, Tuscan, Sardinian, Italian, Basque, French, Russian, Orcadian). The Adyghe is the closest European population to the Jewish population (Kopelman et al., 2009).

Figure 4.3.1: Neighbor-joining tree based on the allele-sharing genetic distance.



CHAPTER FIVE

5. DISCUSSION

The aim of this study was to define ROH segments of a sample of individuals in Palestine. The study proved that individuals of parental relatedness have long ROH segment in their genomes. Our ROH analysis was based on SNP data methods that aimed to define ROH segments among Palestinian population. Such ROH segments contain genes linked with diseases. It has been widely known that ROH segments carry alleles with recessive variants. The results are based on the study of consanguineous marriages from a selected cohort of Palestine. The percentage of consanguineous marriages in the Palestinian population is relatively high, with a rate of 45% to 50% in the years 1995 and 2004 respectively. First cousin marriage prevailed in the aforementioned years, comprising 28% to 30% respectively in these two intervals of all marriages. This traditional marriage is the same as other marriages in the whole Arab countries for the same reasons, namely social and economic reasons. These marriages are considered more stable due to shared and familiar cultural traditions of the same family. The harmful effect of inbreeding increases homozygosity rate in recessive genes as expressed in the offspring's of consanguineous marriages. Many identification techniques of diseased genes are available; the genes and the chromosomes are mapped for homozygous segments for individuals with parental relatedness. The results of this study show a list of genes that were mapped in ROH segments of the Palestinian sample as shown in table 4.2.1. The final step is a comparison between the Palestinian cohort and the world wide populations, to get the shared regions as shown in the table 4.3.1.

CHAPTER SIX

6. CONCLUSION

ROH analysis is used to determine the genomic variations for inferring the history of populations, and it also used for gene mapping to find diseases linked with genes. In this study, ROH analysis of samples from the Palestinian population is considered satisfactory indicators for recent inbreeding, where ROH length>=1.5MB. Genes that are located in these ROH segments contain deleterious variants; this suggests that inbreeding is a main reason for homozygosity of delirious variants.

7. REFERENCES

Alvarez, G., Quinteiro, C. and Ceballos, F.C. (2011) 'Inbreeding and Genetic Disorder', in Ikehara, K. (ed.) *Advances in the Study of Genetic Disorders*, InTech.

Assaf, S., Khawaja, M., Dejong, J., Mahfoud, Z. and Yunis, K. (2009) 'Consanguinity and reproductive wastage in the Palestinian territories. Paediatr Perinat Epidemiol', *Researchgate*.

Bennett, R. (2016) *Consanguinity Fact Sheet*, [Online], Available: http://www.larasig.com/node/2020 [6 April 2016].

Bennett, R.L., Motulsky, A.G., Bittles, A., Hudgins, L., Uhrich, S., Doyle, D.L., Silvey, K., Scott, C.R., Cheng, E., McGillivray, B., Steiner, R.D. and Olson, D. (2002) 'Genetic Counseling and Screening of Consanguineous Couples and Their Offspring: Recommendations of the National Society of Genetic Counselors', *Journal of Genetic Counseling*.

Berry, D.P., McParland, S., Bradley, D.G. and Purfield, D.C. (2012) 'Runs of homozygosity and population history in cattle', *BMC Genetics*.

Bittles, A. (2003) 'Consanguineous marriage and childhood health. Developmental Medicine & Child Neurology', *Developmental Medicine & Child Neurology*.

Botstein, D. and Lander, E.S. (1987) 'Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children', *Science*.

Broman, K.W. and Weber, J.L. (1999) 'Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain', *The American Journal of Human Genetics*.

Carothers, A.D., Rudan, I., Kolcic, I., Polasek, O., Hayward, C., Wright, A.F., Campbell, H., Teague, P., Hastie, N.D. and Weber, J.L. (2006) 'Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches', *Annals of Human Genetics*.

Charlesworth, B. and Charlesworth, D. (1999) 'The genetic basis of inbreeding depression', *Pubmed*.

Charlesworth, B. and Hughest, K. (1996) 'Age-specific inbreeding depression and components of genetic variance in relation to the evolution of senescence', *Proc Natl Acad Sci U S A*.

Charlesworth, D. and Willis, J.H. (2009) 'The genetics of inbreeding depression', *Nature Reviews Genetics*.

Christofidou, P., Nelson, C.P., Nikpay, M., Qu, L., Li, M., Loley, C., Debiec, R., Braund, P.S., Denniff, M., Charchar, F.J., Arjo, A.R., Tre´goue¨t, D.A., Goodall, A.H., Cambien, F., Ouwehand, W.H., Roberts, R., Schunkert, H., Hengstenberg, C., Reilly, M.P., Erdmann, J. et al. (2015) 'Runs of Homozygosity: Association with Coronary Artery Disease and Gene Expression in Monocytes and Macrophages', *The American Journal of Human Genetics*.

Gamsiz, E.D., Viscidi, E.W., Frederick, A.M., Nagpal, S., Sanders, S.J., Murtha, M.T., Schmidt, M., Triche, E.W., Geschwind, D.H., State, M.W., Istrail, S., Cook Jr, E.H., Devlin, B. and Morrow, E.M. (2013) 'Intellectual Disability Is Associated with Increased Runs of Homozygosity in Simplex Autism', *The American Society of Human Genetics*.

Gibson, J., Morton, N.E. and Collins, A. (2006) 'Extended tracts of homozygosity in outbred', *Human Molecular Genetics*.

Golden Helix, I. (2016) *Runs of Homozygosity Analysis*, [Online], Available: http://doc.goldenhelix.com/SVS/latest/svsmanual/roh.html.

Hamamy, H. (2012) 'Consanguineous marriages', *J Community Genet*.

Hartl, D.L. and Clark, A.G. (2007) *principles of population genetics*, 4th edition.

Hill, W.G. and Mackay, T.F.C. (2004) 'D. S. Falconer and Introduction to Quantitative Genetics', *Genetics*.

Hosking, F.J., Papaemmanuil, E., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A.E., Allan, J.M., Taylor, M., Tomlinson, I.P., Greaves, M. and Houlston, R.S. (2010) 'Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk', *PubMed*.

Howrigan, D.P., Keller, M.C. and Simonson, M.A. (2011) 'Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms', *BMC Genomics*.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources', *Nature Protocols*.

Ilaria, G., Flavio, F., Faletra, F., Carella, M., Pecile, V., Ferrero, G.B., Biamino, E., Palumbo, P., Palumbo, O., Bosco, P., Romano, C., Belcaro, C., Vozzi, D. and d'Adamo, A.P. (2014) 'Excess of runs of homozygosity is associated with severe cognitive impairment in intellectual disability', *PubMed*.

Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) 'Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex', *Nature Genetics*.

Joshi, P.K.G.S.F.A..T..M.H..N.G.I..J.A.U..T..C.P.O. (2014) 'Runs of homozygosity reveal inbreeding depression on cognitive function and stature'.

Kefi, R., Hsouna, S., Ben Halim, N., Lasram, K., Romdhane, L., Messai, H. and Abdelhak, S. (2015) 'Phylogeny and genetic structure of Tunisians and their position within Mediterranean populations', *Taylor & Francis*, pp. 593-604.

Keller, M.C., Visscher, P.M. and Goddard, M.E. (2012) 'Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data', *Genetics*.

Kirin, M., McQuillan, R., Franklin, C.S., Campbell, H., McKeigue, P.M. and Wilson, J.F. (2010) 'Genomic Runs of Homozygosity Record Population History and Consanguinity', *PLOS ONE*.

Knapp, B. (1993) *Reproduction and heredity*, Atlantic Europe Pub.

Kopelman, N.M., Stone, L., Wang, C., Gefel, D., Feldman, M.W., Hillel, J. and Rosenberg, N.A. (2009) 'Genomic Microsatellites Identify Shared Jewish Ancestry Intermediate Between Middle Eastern And European Populations', *BMC Genet*.

Kuningas, M., McQuillan, R., Wilson, J.F., Hofman, A., van Duijn, C.M., Uitterlinden, A.G. and Tiemeier, H. (2011) 'Runs of Homozygosity Do Not Influence Survival to Old', *PLOS ONE*.

Lencz, T., Lambert, C.G., DeRosse, P., Burdick, K.E., Morgan, T.V., Kane, J.M., Kucherlapati, R. and Malhotra, A.K. (2007) 'Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia', *PubMed*.

Leutenegger, A.-L., Labalme, A., Ge´nin, E., Toutain, A., Steichen, E., Clerget-Darpoux, F. and Edery, P. (2006) 'Using Genomic Inbreeding Coefficient Estimates for Homozygosity Mapping of Rare Recessive Traits: Application to Taybi-Linder Syndrome', *The American Journal of Human Genetics*.

Magi, A., Tattini, L., Palombo, F., Benelli, M., Gialluisi, A., Giusti, B., Abbate, R., Seri, M., Gensini, G.F., Romeo, G. and Pippucci, T. (2014) 'H3M2: detection of runs of homozygosity from whole-exome sequencing data', *Oxford Journals*.

Mahdi, G., Christiane, R., Rong, C., Narayan Vardarajan, B., Jun, G., Sato, C., Naj, A., Rajbhandary, R., Wang, L.-S., Valladares, O., Lin, C.-F., Larson, E.B., Graff-Radford, N.R., Evans, D., Jager, P.L.D., Crane, P.K., Buxbaum, J.D., Murrell, J.R., Raj, T., Ertekin-Taner, N. et al. (2015) 'Association of Long Runs of Homozygosity With Alzheimer Disease Among African American Individuals', *JAMA Neurol*, pp. 1313-1323.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A.k., Farrington, S.M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S.H., Dunlop, M.G., Wright, A.F., Campbell, H. et al. (2008) 'Runs of homozygosity in European populations', *The American Journal of Human Genetics*.

Medicine, U.o.P.S.o.V. (2005) *Calculating the Coefficient of Inbreeding in Simple Pedigrees*, [Online], Available: http://cal.vet.upenn.edu/projects/genetic/inbreed/coeff/page1.htm [2016].

Miano, M.G., Jacobson, S.G., Carothers, A., Hanson, I., Teague, P., Lovell, J., Cideciyan, A.V., Haider, N., Stone, E.M., Sheffield, V.C. and Wright, A.F. (2000) 'Pitfalls in homozygosity mapping', *The American Journal of Human Genetics*.

NCBI (2016) *HEXA hexosaminidase subunit alpha [Homo sapiens (human)]*, [Online], Available: http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=3073.

Nothnagel, M., Lu, T.T., Kayser, M. and Krawczak, M. (2010) 'Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans', *Human Molecular Genetics*.

Pemberton, T..A.D..F.M..M.R..R.N.a.L.J. (2012) 'Genomic Patterns of Homozygosity in Worldwide Human Populations', *The American Journal of Human Genetics*, pp. 275-292.

Pemberton, T.J. (2015) *Pemberton lab*, [Online], Available: http://pembertonlab.med.umanitoba.ca/index.html.

Pemberton, T.J., Absher, D., Feldman, M.W. and Myers, R.M. (2012) 'Genomic Patterns of Homozygosity in Worldwide Human Populations', *The American Society of Human Genetics*.

Pollinger, J.P., VonHoldt, B.M., Lohmueller, K.E., Han, E., Parker, H.G., Quignon, P., Degenhardt, J.D., Boyko, A.R., Earl, D.A., Auton, A., Reynolds, A., Bryc, K., Brisbin, A., Knowles, J.C., Mosher, D.S., Spady, T.C., Elkahloun, A., Geffen, E., Pilot, M., Jedrzejewski, W. et al. (2010) 'Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication', *Nature*.

Purcell, S. (2010) *PLINK (1.07) Documentation*.

Quinlan, A.R. and Kindlon, N. (2016) *bedtools: a powerful toolset for genome arithmetic*¶, [Online], Available: http://bedtools.readthedocs.io/en/latest/.

Risch, N. and Botstein, D. (2003) 'Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease', *Nature Genetics*, Mar, pp. 228-237.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., Schaffner, S.F. and Lander, E.S. (2007) 'Genome-wide detection and characterization of positive selection in human populations', *Nature*.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapa, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., Schaffner, S.F. and Lander, E.S. (2007) 'Genome-wide detection and characterization of positive', *Europe PMC Funders Group*.

Shahin, H., Walsh, T., Abu Rayyan, A., Lee, M.K., Higgins, J., Dicke, D., Lewis, K., Thompson, J., Baker, C., Nord, A.S., Stray, S., Gurwitz, D., Avraham, K.B., King, M.-C. and Kanaan, M. (2010) 'Five novel loci for inherited hearing loss mapped by SNP-based homozygosity profiles in Palestinian families', *European Journal of Human Genetics*.

Spain, S.L., Cazier, J.-B., Houlston, R., Carvajal-Carmona, L. and Tomlinson, I. (2009) 'Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom', *PubMed*.

Szpiech, Z.A., Xu, J., Pemberton, T.J., Peng, W., Rosenberg, N.A., Li, J.Z. and Zöllner, S. (2013) 'Long Runs of Homozygosity Are Enriched for Deleterious Variation', *The American Society of Human Genetics*.

Vine, A.E., McQuillin, A., Bass, N.J., Pereira, A., Kandaswamy, R., Robinson, M., Lawrence, J., Anjorin, A., Sklar, P. and Gurling, H.M.D. (2009) 'No evidence for excess runs of homozygosity in bipolar disorder', *Psychiatric Genetics*.

Waller, D.M. and Keller, L.F. (2002) 'Inbreeding effects in wild populations', *Elsevier Science*.

Wright, S. (1922) 'Coefficients of Inbreeding and Relationship', *The American Naturalist*.

Wright, S.I. and Andolfatto, P. (2008) 'The Impact of Natural Selection on the Genome: Emerging Patterns in Drosophila and Arabidopsis', *Annual Reviews*.