Genome Wide Association Studies for Identification of Candidate Risk Factors for Juvenile Onset of Demodicosis Disease

By

Qurrat ul Ain

NUST201463259MRCMS64014F

A thesis submitted in partial fulfillment of the requirement for the degree of

Master of Science

In

Computational Sciences & Engineering

Research Center for Modeling and Simulations

(RCMS)

National University of Sciences and Technology (NUST), Pakistan

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: **Qurrat ul Ain** Reg no: **NUST201463259MRCMS64014F**

Titled: **Genome wide association studies for identification of candidate risk factors for juvenile onset of demodicosis disease** be accepted in partial fulfillment of the requirements for the award of **MS CS & E** degree with (_____________Grade).

Examination Committee Members

1. Name: Dr. Mehak Rafiq Signature:_______________

2. Name: Dr. Rabia Amir Signature:_______________

3. Name: Dr. Zamir Hussain Signature:_______________

Supervisor's name: Dr. Shumaila Sayyab Signature:_______________

Date:___________________

___________________ ____________

Head of Department Date

COUNTERSIGNED

Date: _______________ ________________

Dean/Principal

Declaration

I hereby declare that this thesis comprises of my own research work; any part of this thesis is not plagiarized. The contributions of different people in the form of suggestions, discussions and previously published literature are acknowledged and duly referenced

Qurrat ul Ain

NUST201463259MRCMS64014F

DEDICATED TO

MY DEAR PARENTS, HUSBAND & MY SONS

Acknowledgements

All the praises and gratitude be to Allah, who bestowed upon me the ability to accomplish my research project successfully. The thesis was started with an objective to identify the candidate risk factors for juvenile onset of canine demodicosis via Genome Wide Association Studies. To achieve this objective was almost impossible without timely guidance, scholarly supervision and encouragement of my honorable supervisor Dr. Shumaila Sayyab, Assistant Professor Research Center for Modeling and Simulation (RCMS), National University of Science and Technology (NUST). Her encouragement and guidance to keep me on the track to achieve my objective is commendable, memorable and incomparable for me.

I am grateful to the Guidance and Examination Committee members Dr. Zamir Hussain RCMS, NUST; Dr. Mehak Rafiq RCMS, NUST and Dr. Rabia Amir ASAB, NUST for providing me guidance to complete this seemingly formidable job.

I acknowledge my colleagues i.e. Mehwish Noureen, Maria Altaf Satti, Fauzia Ehsan and Tabeer Fatima for the cognitive and moral support, they provided to me. I am also thankful to Sohaib Aslam for his guidance to use ANNOVAR.

I appreciate the efforts of my parents, brothers and my husband who all stood by me as and when required to overcome any hardships appeared in completion of this project. Most importantly I am grateful to my sons, who have sacrificed many golden moments of their lives because of the busy routine of their mother.

Qurrat ul ain

Abstract

Genome wide association Studies (GWAS) are an efficient approach to estimate the candidate risk loci associated with complex diseases. Juvenile onset of canine demodicosis is a common inflammatory disease of the skin of dogs. It is severely invasive and even fatal in some cases. Suppression of immune response as an underlying cause of infestation of demodicosis is well established. Genetic causes of the disease are yet unknown. Therefore, this thesis focuses on identifying the candidate single nucleotide polymorphic risk loci associated with juvenile onset of canine demodicosis through GWAS. Ten candidate SNPs were successfully identified to be significantly associated with the disease in a discovery phase. All these SNPs are located in intergenic region on chromosome 28. Literature search showed that all four genes neighboring these significant SNPs are directly or indirectly involved in inflammatory related diseases and with skin and immune system as related phenotypes in other species like humans. Therefore, we suggest that these genes might be good candidates for future research, to identify the causal genetic abnormalities of the disease. In further applications, this study can provide new dimensions in diagnostic and treatment domains related to demodicosis in dogs and once established in dogs as model organism, it can further be extended to benefit humans.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1. Introduction

The chapter includes introduction of genome wide association studies (GWAS) and juvenile onset of canine demodicosis.

1.1 Genome wide association studies

GWAS is an approach that helps to locate differences in nucleotide sequences throughout the genomes. This is done in order to identify susceptible nucleotide loci for generally prevailing complex diseases among populations. It helps to prognosticate the persons to be at risk in future and to develop prevention and cure of the disease [1]. Most of the common diseases are the result of genomic mutations (variations) in more than one gene and are therefore called as complex diseases [1]. Though the effect size of these genes may be too low, yet they together with the environmental conditions lead to disease [2, 3]. The objective of the GWAS is to determine not only the causal variants but also know how these variants contribute to the disease etiology and to find regulatory non-coding mechanisms involved [4]. Single nucleotide polymorphs (SNPs) are mapped throughout the genomes to find those significantly associated with a given disease [5].

1.1.1 Significance of GWAS

GWAS have significantly performed the task of identifying the causal variants in coronary artery disease [6], blood pressure [7], hepatitis B and hepatic cancer [8], human height [9], type II diabetes [10]. It has also helped in identification of risk factors of neurological disorders like attention deficit hyperactivity (ADHD), autism, bipolar disorder, major depressive disorder, and schizophrenia [11]. To gain insight into heritability of complex diseases and foretell the chances of occurrences of common diseases has been achieved in many cases through determination of various single polymorphic loci associated with the diseases after the advent of GWAS [12].

The very first study on GWAS proved the presence of polymorphic alleles in *CFH* gene coding for complement factor H protein, resulting in a change of amino acid from tyrosine to histidine. This change proved to be associated with age related macular degeneration and hereby blindness in old age [13]. GWAS is able to identify variants linked with the metabolism and efficacy of drugs as in case of warfarin. Many genes have been located through identification of variants that influence the doses of warfarin. As a consequence, patient related treatment of the doses to prevent toxic effects of anticlotting drug is recommended. Therefore, genetic tests of patients are proposed before warfarin dosing to avoid these adverse effects. It is a step forward to personalized medicine [1]. The National Human Genome Research institute has mentioned 1818 GWAS identifying 12498 risk loci until March 2014 [14]. Till 2015 this number has been increased up to 5582 as reported by GWAS catalog [15] as shown in **figure 1.1**.

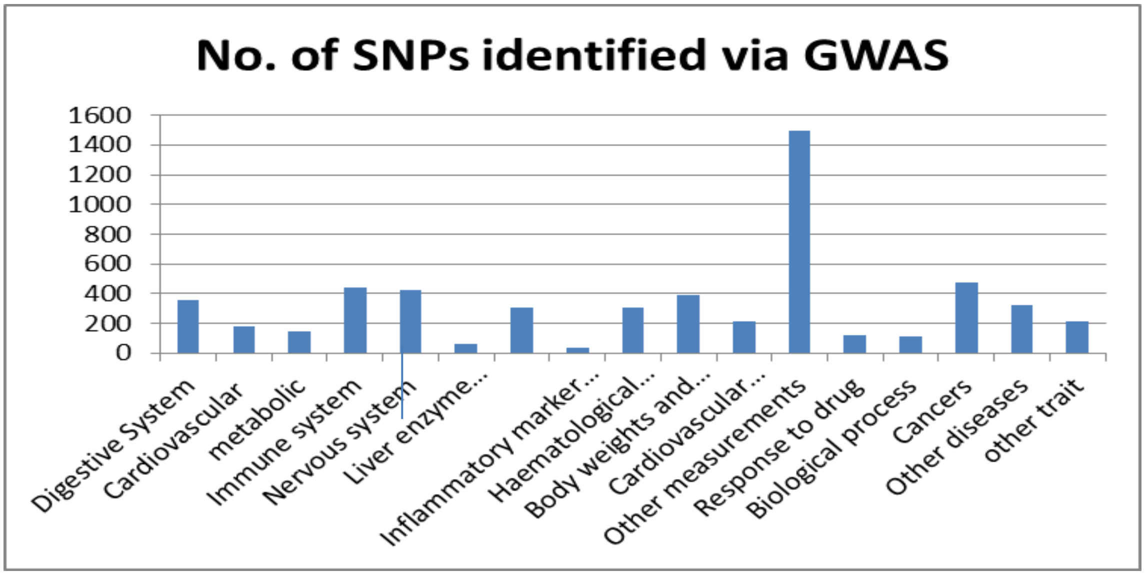


Figure 1.1: Total 5582 SNPs reported in GWAS Catalog

1.1.2 Single nucleotide polymorphims (The markers)

Single nucleotide polymorphisms (SNPs) are said to be the variations responsible for etiology of complex diseases in many cases [16]. SNPs constitute complex genetic architecture of the disease [17]. SNPs differ in population at a single base pair position with a proportion of 1 in every 100 individuals [16]. In most cases SNPs occur in two out of four possible polymorphic conditions as shown in figure 1.2. This bi-allelic condition leads to the feasibility of their mapping through high throughput microarray technology [5]. GWAS approach looks for up to billions of SNPs throughout the organism's genome using mapping of tag SNPs [18].

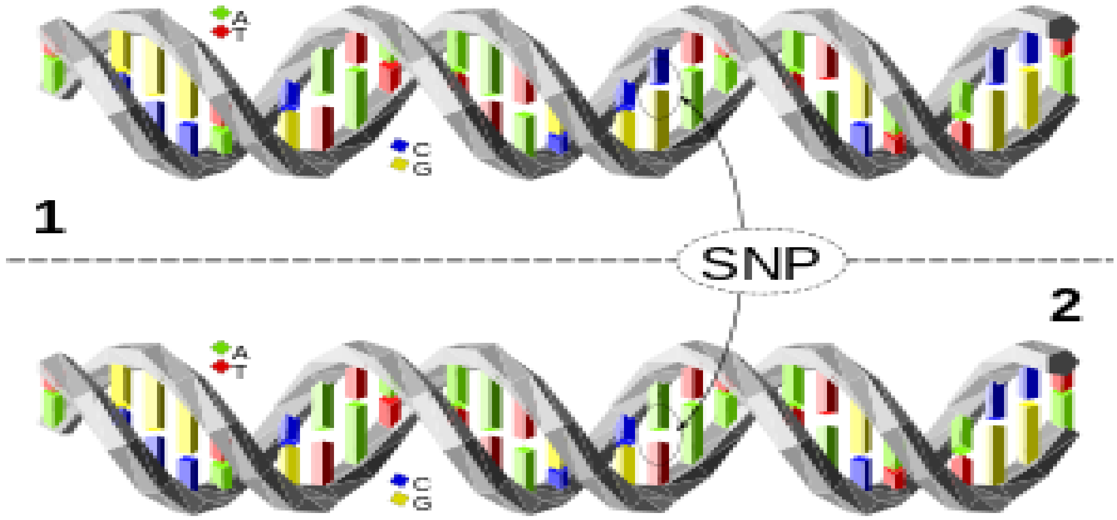


Figure 1.2: A SNP having CG base pair in one person and AT base pair in another person at same position in another person

The foundation for GWAS is laid on following developments in the field of genomics which are as follow; sequenced genomes [19], identification and verification of millions of common variants present across the population genomes in a frequency greater than 5%. Selection of tag SNPs via the phenomenon of linkage disequilibrium to obtain desired coverage became possible after the identification of LD structures across the genomes of different populations through International Hap Map consortia [20, 21]. Microarray genotyping chips to assay these variants have also been developed, to gain maximum through put for GWAS. Large sample size is also a prerequisite to identify significant variants, associated with the phenotype of interest [22].

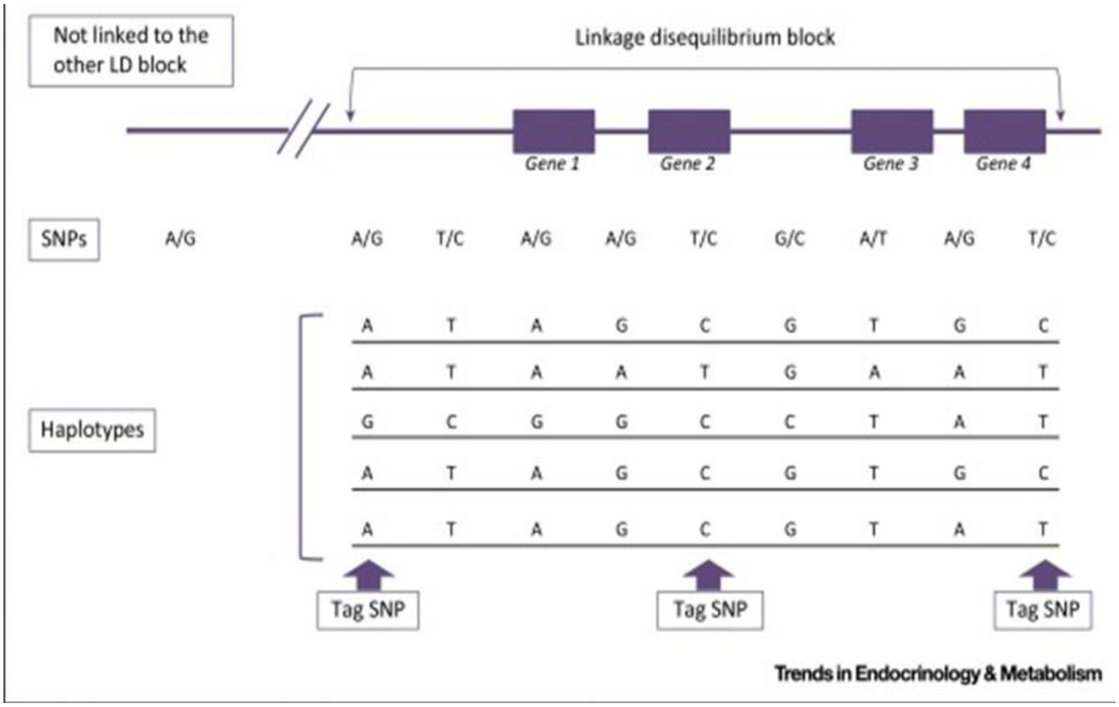
1.1.3 Assumptions for GWAS

1.1.3.1 Common disease common variant hypothesis

One of the assumptions underlying GWAS is the common disease common variant hypothesis (CD/CV) which states that same susceptible alleles are present in significant numbers among the patients sharing same disease [23].

1.1.3.2 Linkage disequilibrium

Most of the SNPs in the genome are associated through the phenomenon of linkage disequilibrium (LD). They are so intimately located on same chromosome that they are inherited together [21], until separated by genetic recombination whose chances are very rare in nature [16]. Since these recombination events occur on particular sites of the chromosome called as hot spots for the genetic interchange, [24] the linked regions constitute a haplotype block **(Figure 1.3)** [21].



Trends in Endocrinology & Metabolism

Figure 1.3: Four genes present in linkage disequilibrium with each other and hence forming a haplotype block. Tag SNPs are selected at appropriate distance within a single haplotype block to achieve maximum coverage at optimum genotyping cost.

The success and accelerated rate of the GWAS have become possible after the completion of “international Hap Map project”. It was aimed to map the regions of single nucleotide polymorphisms (SNPs) inherited together because of linkage disequilibrium [25]. International Hap Map project enabled to extract a comprehensive collection of SNPs termed as “Tag SNPs’’. These “Tag SNPs” can be used as markers to map linked

SNPs in the same haploblocks of individual genomes, to identify SNPs associated with the complex diseases [26]. Through DNA (deoxyribonucleic acids) microarray technology, chips containing millions of these tag SNPs are produced on commercial bases. These genotyping chips are able to assess all the SNPs present in that chip across the genomes of genotyped individuals (cases and controls) and hence, check for the variants at particular loci without any prior knowledge [27]. Human genome bears 7.5 million SNPs with minor allele frequency > 0.05; only 100,000 are typed on microarray genotyping chips [28]. However, most of the SNPs that are not typed, yet can be associated to the phenotype, are detected because they correlate due to linkage disequilibrium [29]. They occur in such proximity to tag SNPs on the same chromosome that they are always inherited together [29]. Linkage is measured through relatively common regions of chromosomes among related individuals, which are called as haplotypes and are said to be inherited by descent [30].

1.1.3.3 Population stratification

Differences found in allele count within the members of a study group due to differences in ancestry are termed as population stratification. Population stratification, if not corrected, results in false positive associations [31]. Diverse populations differ in the respective frequencies of polymorphic loci. Likewise there is diversity in presence or absence of a particular locus to be susceptible in different populations [32]. In contrast to population stratification, cryptic relatedness causes spurious associations because of the presence of too close ancestors.

1.1.3.4 Selection of tag SNPs

Selection of appropriate SNP markers to map through the whole genome with best possible coverage (tag SNPs), is another pre-requisite of GWAS for precision of results and economical concerns of genotyping **(Figure 1.4).**

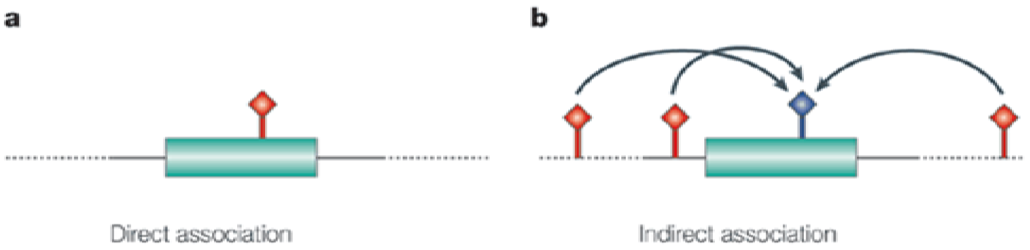


Figure 1.4: Tag SNPs directly or indirectly associate to the phenotype of interest

Linkage disequilibrium enables the mapping of tag SNPs across the genomes of samples. If a tag SNP is not in a haplotype block of the associated SNPs that portion of block is left unmapped leading to loss of associated alleles in a particular study. Economical tag SNP selection is possible because of linkage disequilibrium as 500,000 tag SNPs are sufficient to map the whole human genome [33].

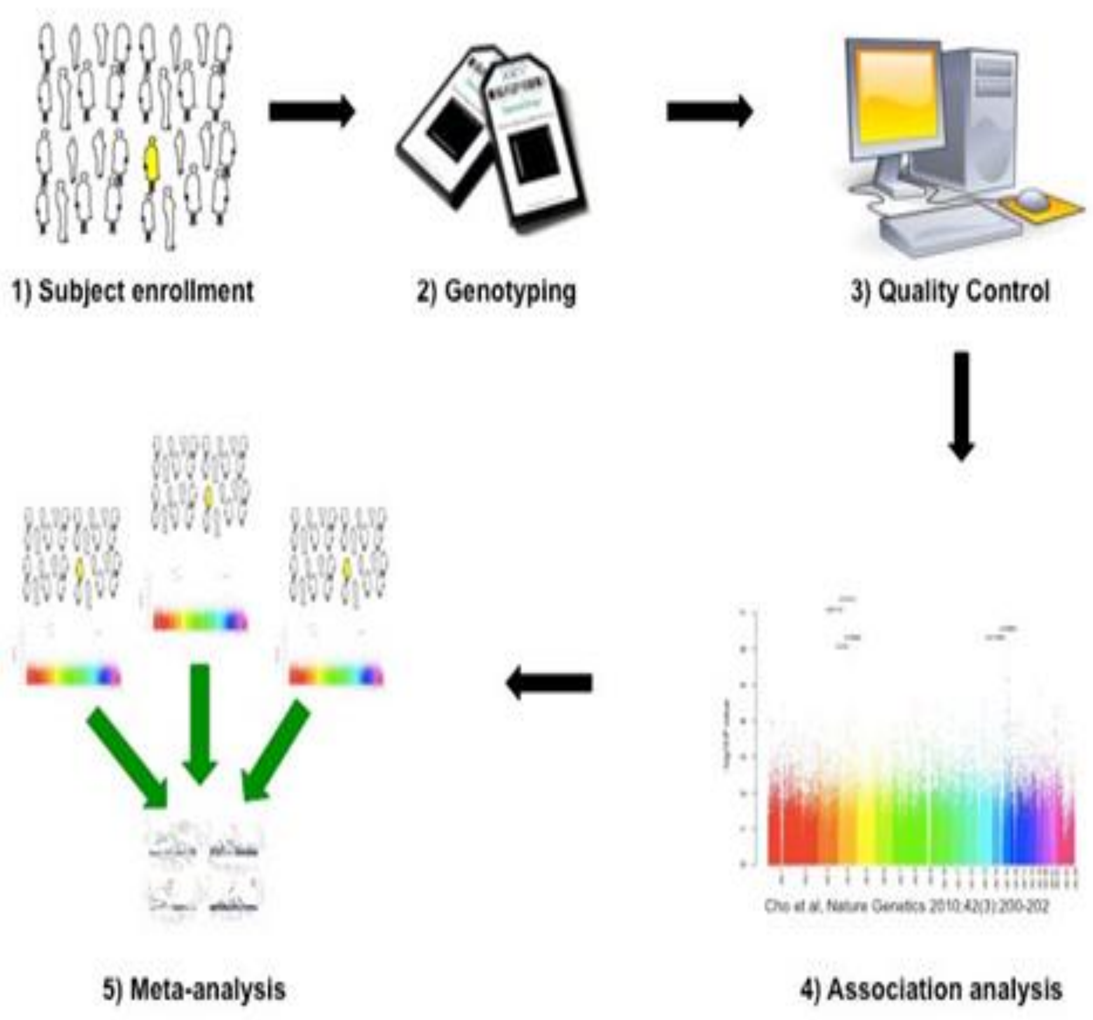
1.1.3.5 Sample size

Large sample size (usually in thousands) of cases and controls is the prerequisite of the GWAS in humans [34] to determine those variants predisposing to the diseases like obesity, coronary artery disease, type II diabetes. Whereas, in case of dogs, small sample size (20 for mammalian and 100 cases and 100 controls for complex diseases) are required due to larger blocks of LD, as described in a study by Karlsson and Linblad-Toh in 2008.

1.1.4 Steps and concerns of genome wide association studies

1.1.4.1 Subject selection

Patients suffering from a particular disease are chosen as cases (**Figure 1.5)**. In contrast, people who do not have any of the symptoms of that disease are selected as controls. The only condition is that subjects chosen as controls should represent the same population as cases. In other words cases and controls in a case-control GWAS study should form a group with respect to age, sex and race to avoid the spurious associations caused by population stratification. A study reported the use of 3000 common controls for every 2000 cases of each of the seven diseases under study and it had least effects on genotype distribution. All the subjects were UK (United Kingdom) nationals [34]. Taking common controls for several phenotypes results in loss of power, but the issue can be resolved by large sample size selection. Issues of cryptic relatedness can also result in inflation of test statistic but it can be overcome by various approaches like genomic control [36].



1) Subject enrollment

2) Genotyping

3) Quality Control

4) Association analysis

5) Meta-analysis

Cho et al, Nature Genetics 2010;42(3):200-202

Figure 1.5: Pipeline of GWAS. Steps 1-4 are followed in an individual GWA Study. Meta-analysis is a combined analysis of different independent GWAS studies to reach at a conclusion regarding association of a SNP to the disease.

1.1.4.2 Genotyping

Genome wide typing of the tag SNPs across the genomes of all the samples of cases and controls to search for SNPs associated with the phenotype of interest is crucial step of

GWAS [37]. Two commercial platforms of genotyping are famous to efficiently perform this purpose. These are Affymatrix and Illumina [37]. Present study includes samples genotyped with Illumina HD Canine SNP array. It is a high density genotyping chip covering all the SNPs in the genome.

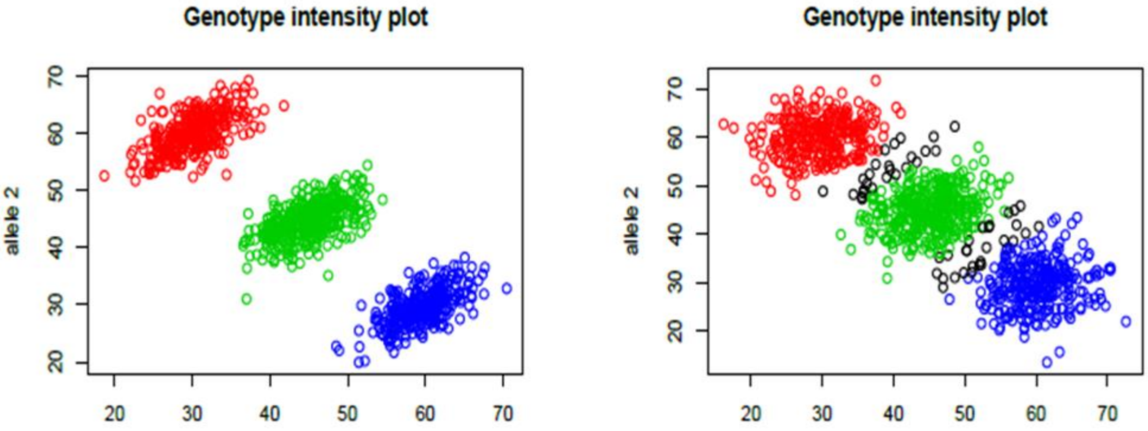
1.1.4.2.1 Sample preparation for genotyping

After extraction of DNA, it is prepared for hybridization which is then performed through array based genotyping chips. Two platforms i.e. illumina and affymatrix are renowned for commercial production of genotyping chips based on genomic knowledge regarding the selection of markers to attain maximum possible coverage of genome or genomic regions. When hybridization step is completed the arrays are cleaned and a raw file is generated. The raw file named as .dat file consists of optical images of hybridization probes. After getting the optical images the strength of their intensities is computed through pixel values and is stored in another file named as .cel file. Each signal intensity represents each cell of the image and which indicates the data for each subject of the study. Signal intensities of these .cel files are normalized and following this step genotypes are obtained from these signal intensities [37].

1.1.4.2.2 Genotype calling

An allele is said to be homozygous if its signal intensities are higher than the other allele. In contrast, if the intensities are equal for both the alleles of a SNP then they are said to be heterozygous. In this way genotypes are assigned to all the SNPs present on the array chips. Intensive computational effort is required which is complemented by efficient genotype calling algorithms. Therefore, there is possibility of three types of genotypes for each tag SNP. It would be either homozygous for any of the two alleles (AA or BB) or it would be heterozygous (AB) for a particular tag SNP. All the subjects are hence genotyped for all the tag SNPs of the microarray genotyping chip.

Hence, the genotype data for each SNP should appear to be in three clusters of different colors representing any one of the three genotypes. In case sufficient number of genotype is not assigned to a particular SNP, it is termed as having missing genotypes and is excluded from the analysis. In other words, if the SNP is not genotyped in many of the subjects included in the study (black dots in figure 1.6), it is not of any use and hence is discarded (**Figure 1.6**) [37].



1.6 A (Genotype calling) 1.6 B

Clusters well defined Overlapping in clusters

Colors description: red (AA), green (Aa), blue (aa) and black for no call

1.1.4.2.2.1 Minor allele frequency (MAF)

In such a situation, when abundant genotypes of a particular SNP are monomorphic genotypes for any one of the possible homozygous genotypes either AA or BB. Consequently, it is impossible to discriminate minor allele and major allele. Such alleles are of no use in GWA studies because calculation of association statistics is based on the minor allele frequency which is not obtained through monomorphic SNPs. This is the reason why minor allele frequency threshold is set in quality control analysis and the SNPs failed therein are discarded [37].

1.1.4.2.2.2 Hardy Weinberg Equilibrium (HWE)

The principal on which all the population genetic studies rely is Hardy Weinberg Equilibrium (HWE) which states that there is a tendency of constancy for allele and genotypic frequencies in the population provided that following assumptions are met;

large population size, random fertilization, no migration, no selection, fixed relationship between alleles and their respective genotypes (**Figure 1.7**) [37].

An interpretation of fixed relationship between allele and genotypes in the perspective of quality control is that genotyping errors may lead to deviation from HWE. So SNPs that are too much out of HWE may be a consequence of artifacts of genotyping. Choice of a disease status also violates the fourth assumption of HWE so; the SNPs that are too much out of HWE are also discarded [35].

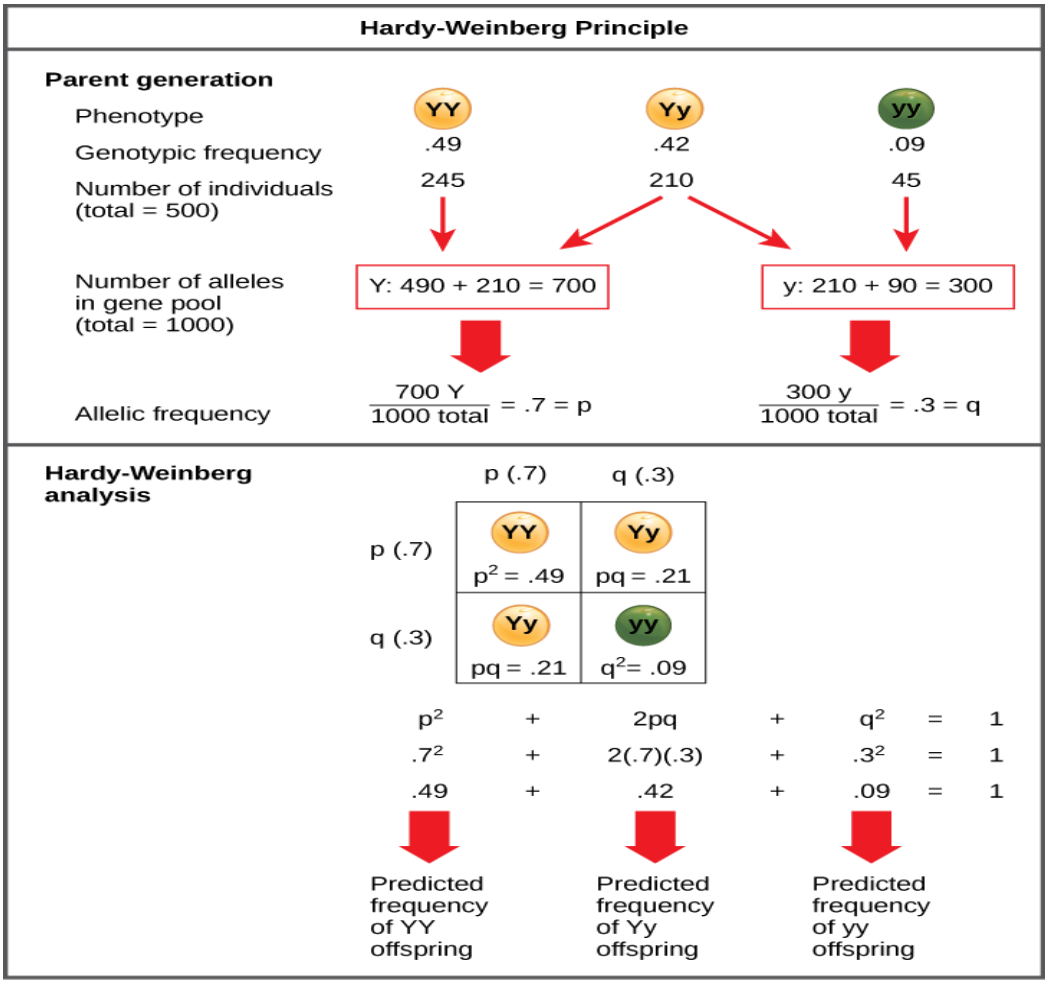


Figure 1.7: an illustration of hardy Weinberg equilibrium

1.1.4.2.3 Subjects calling for genotypes

Subjects genotyped for less than 97% SNPs are considered as having missing SNPs and therefore should be removed from the analysis. It is necessary for powerful case control analysis of the trait of interest in a particular GWA study. The threshold for individual call rate for SNPs may also be increased from (>3% to > 5%) or even >10% as recommended by the genotyping contract authorities [37].

Heterozygosity is a good indicator of the quality of subjects. A subject containing too much heterozygosity may be considered a poor sample because of the manual artifacts such as DNA contamination. So mean and Standard deviation of heterozygosity is computed across all the subjects. Those subjects being out of range (M $\pm 3SD$) are eliminated from the data [37].

Relatedness is another criterion to judge quality of two or more samples. For any two of the subjects, if the probability of the SNPs to be identical is more than 50%, the samples are considered to be close relatives. That pair or one of the members of the pair is suggested to be removed from the analysis [37].

1.1.4.2.4 Quality control

Quality control measures are important to make the data homogeneous for analysis. Steps of quality control are as follow [38].

1. Placement of allelic strands to be genotyped to the references on the chip. Forward placement is the usual setting [38].
2. Check for sample relatedness which would lead to inappropriate risk alleles' identification, because of shared ancestry [39].
3. Duplication is also checked and corrected which may otherwise lead to inflation of risk allele identification, if cases are over duplicated. Possibility is an increase in noise signal, if controls are over duplicated [38].
4. SNPs are checked for good quality call rates, minor allele frequency (MAF) and Hardy Weinberg Equilibrium [32]. SNPs having low call rates are eliminated form analysis.
5. Hardy Weinberg equilibrium (HWE) is controlled through prior checking of the SNPs that violate HWE for elimination. SNPs associated to the disease violate HWE therefore,

they must not be eliminated. Therefore optimum threshold for HWE is set very carefully according to situational requirements [37].

6. SNPs highly linked through linkage disequilibrium are also eliminated because of the possibility of sample relatedness [40].

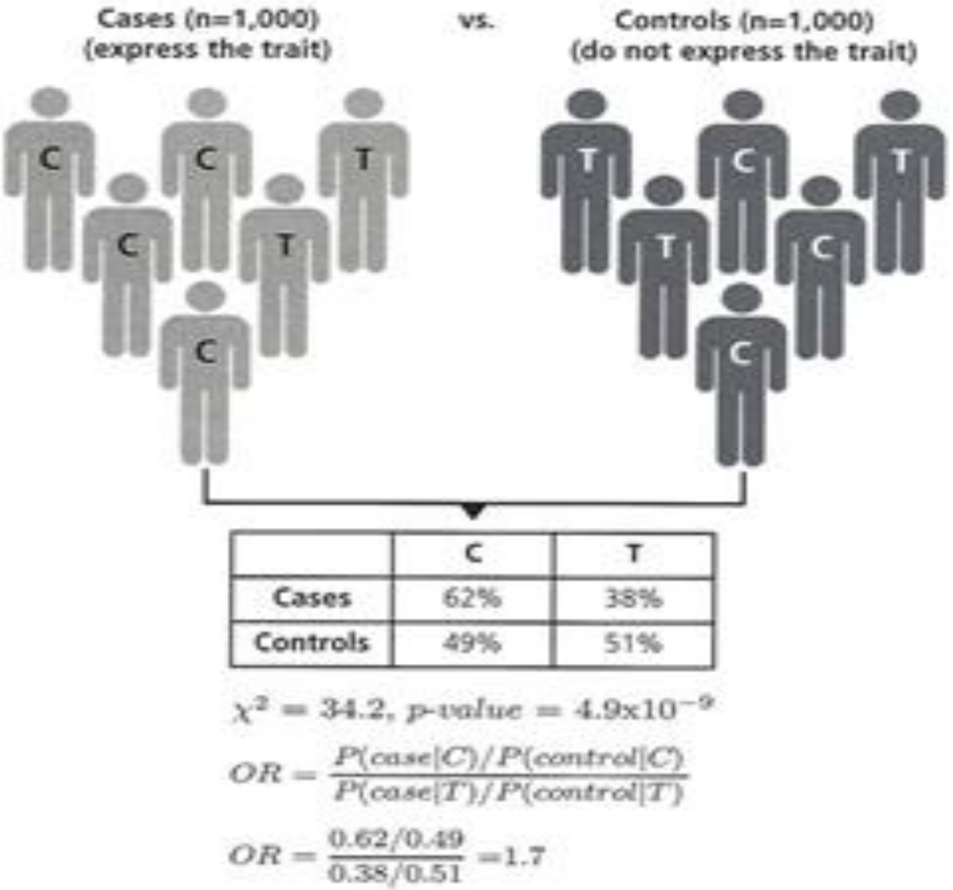
7. Large samples are divided into batches for facilitation in analysis so batch effects must be checked [35, 36]. Batch effects are removed through variables critical for the study being randomized e.g. presence or absence of phenotype, age, sex, BMI (body mass index) [23].

1.1.4.4 Statistical analysis

1.1.4.4.1 Single base-pair position association analysis

Analysis of single base-pair position association to the disease is pivotal statistical test. It analyses the association of MAF of each genotyped SNP individually to the phenotype of interest. Chi-square test and Fisher's exact test are two tests statistics involving contingency table of case and control count of the subjects in rows and genotypes/alleles in the columns as the case may be **(Figure 1.8)**.

Association is measured in most of the GWAS analyses through trend test across the minor allele frequency (MAF) for each SNP. Odds ratio is determined for analysis of individual samples obtained by dividing the presence of the phenotype in the study samples through their presence in background population. Selection of p-value significant enough to find associations of SNPs with the trait of interest is of prime importance. This is done such that true positive SNPs can be identified and false positive SNPs can be removed. Interactions of more than one SNP, their combine role and haplotypes can also be analyzed by GWAS [41].



	C	T
Cases	62%	38%
Controls	49%	51%

$\chi^2 = 34.2$, $p\text{-}value = 4.9\text{x}10^{-9}$

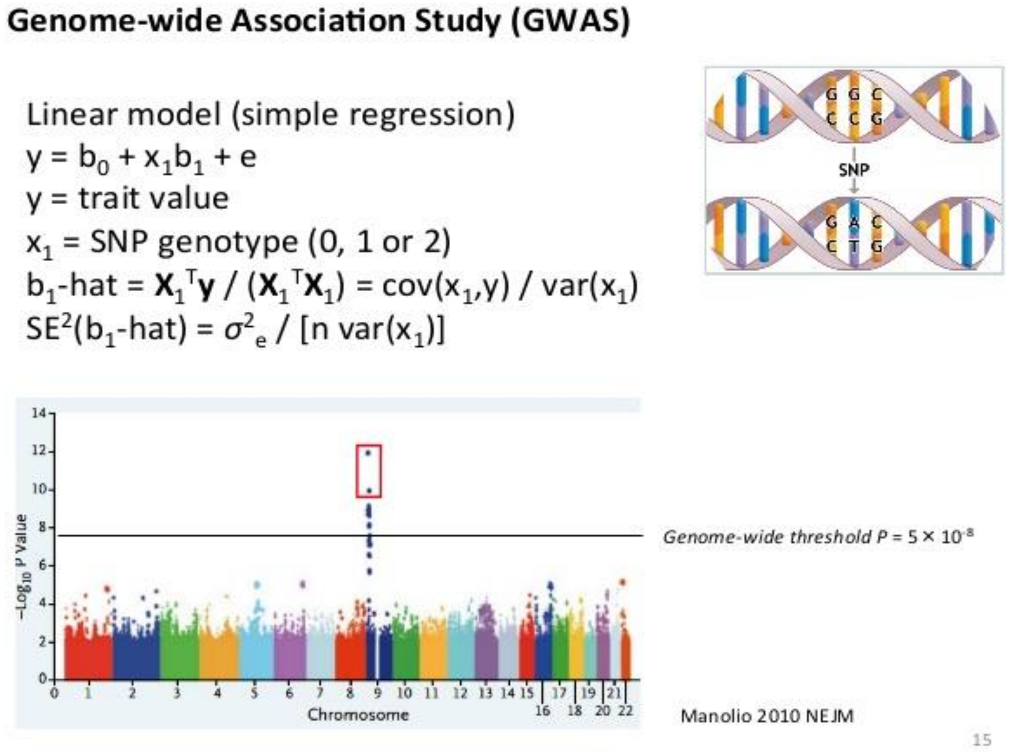
$$OR = \frac{P(case|C)/P(control|C)}{P(case|T)/P(control|T)}$$

$$OR = \frac{0.62/0.49}{0.38/0.51} = 1.7$$

Figure 1.8: Chi-square statistics implied in different tests such as Trend test etc, to find the association of a binary categorical variable with the phenotype of interest which is a disease in many cases.

1.1.4.4.2 Multiple testing analysis

Genotyping of millions of SNPs in thousands of samples of cases and controls requires tens of millions of tests of associations which may cause type 1 error (false positive). One of the ways to remove this error is the Bonferroni correction. Bonferroni correction is used to transform p-values to a threshold value obtained by dividing the p-values to the total SNPs multiplied by total samples [28]. Bonferroni correction is more conservative because of the degree of correlation among tag SNPs [41]. False discovery rate is also calculated to determine how strongly significant SNPs are associated to the disease [42, 43, 44]. Logistic regression including lasso penalized regression is another strategy utilized successfully in cases where predictors are far more than observations (**Figure1.9**) [29].



Genome-wide Association Study (GWAS)

Linear model (simple regression)

$y = b_0 + x_1b_1 + e$

y = trait value

x_1 = SNP genotype (0, 1 or 2)

$b_1\text{-hat} = \mathbf{X}_1^T\mathbf{y} / (\mathbf{X}_1^T\mathbf{X}_1) = \mathrm{cov}(x_1,y) / \mathrm{var}(x_1)$

$SE^2(b_1\text{-hat}) = \sigma^2_e / [n\ \mathrm{var}(x_1)]$

Manolio 2010 NEJM

Figure 1.9: Regression to incorporate the covariates in association analysis

To further avoid the false positives, replication studies can be conducted for exact determination of susceptible variant loci [45]. When individual SNPs are screened for association through multiple testing, suggested p-value threshold is 5*10e-8[46].

1.1.5 Permutation testing

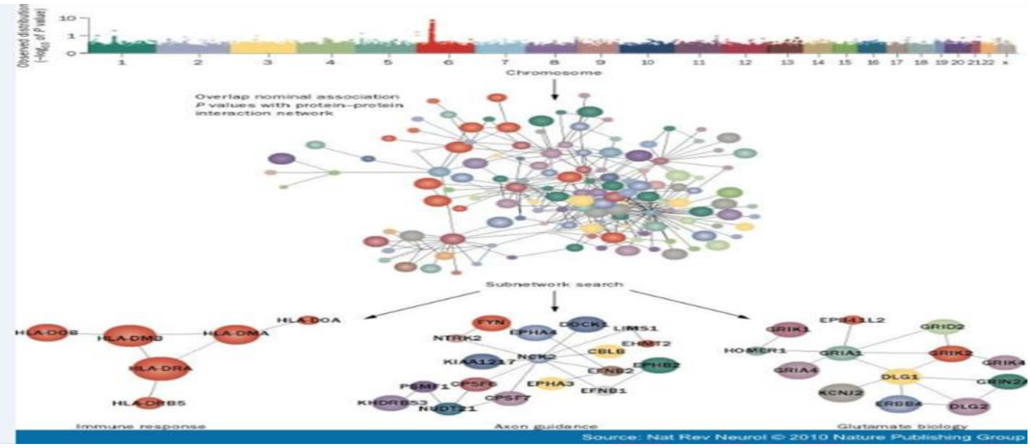
Permutation testing is another technique used to cope up with strong correlation among variables [47]. One of the most important causes of strong correlation might be due to relatedness which leads to population substructure [48]. If correlation is not incorporated while performing statistical analysis, it may result in inflation of test statistics which is an indicator of false positives. To control for population substructure, genomic control (GC) is recommended provided that a complete knowledge of genetic markers is available [48].

1.1.6 Replication

Replication studies are continuation of the first phase called as discovery phase. The discovery phase constitutes the mapping of tag SNPs across the genomes of cases and controls, to screen for the SNPs appearing to be associated with the disease. The second phase called as replication, comprises of another independent case and control cohort of genotyping the SNPs selected in first study, to be verified as the SNPs significantly associated with the disease under study [45].

1.1.7 GWAS pathway analysis

It is a very challenging task to identify actual causal variants through GWAS, because sometimes the associated SNPs may be in strong linkage disequilibrium with the causal SNPs. Most of the associated SNPs are found in the non-coding regions and identifying their role in regulatory mechanisms, is another challenge. The validation of SNPs to be associated to disease can be done through the pathway analysis of related genes. Thus the significant SNPs are subjected to gene annotation analysis followed by pathway analysis (**Figure 1.10**) [31, 39 and 50]. In this way the role genes related to the SNPs identified through GWAS is determined in terms of disease etiology.



Source: Nat Rev Neurol © 2010 Nature Publishing Group

Figure 1.10: GWAS Pathway Analysis

1.1.8 Limitations of GWAS

One of the limitations of GWAS is its inability to detect rare novel variants associated with the disease that are thought to satisfy the missing heritability of the common diseases [39, 31]. Violation of assumption of independence of SNPs for the tests of associations is also another limitation, because the study is based on the phenomenon of linkage disequilibrium of SNPs to their haplotypes. SNPs explored through GWAS, as risk loci for complex diseases, possess too little risk estimate to conduct comparative studies across multiethnic populations. This may lead to false positive estimates due to population specificity. SNPs are more common in populations and they may also share them evolutionarily [51].

1.2 Demodicosis in dogs

Demodicosis is a commonly occurring skin disease in mammals. It is caused by proliferated population of Demodex mites **(Figure 1.11).** More than 50 species of mammals host in different species of Demodex. Even four or more species can be observed in some mammalian species [52] according to different habitat conditions [56].

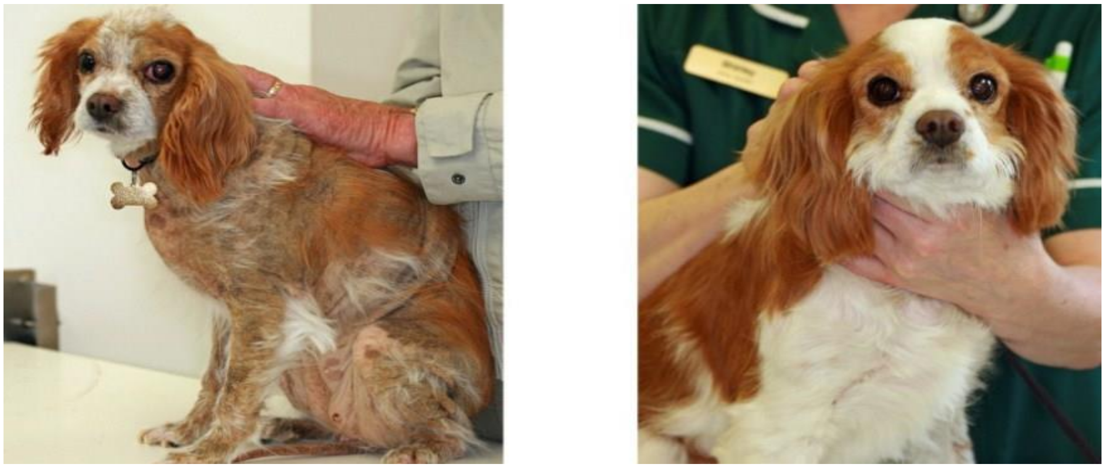


Figure 1.11 A **Canine demodicosis** **Figure1.11B**

A: Demodicosis infestation in a dog B: Same dog before the onset of demodicosis

Demodex is an arthropod that belongs to the family democidae of class arachnida. Canis species was commonly observed in dogs two centuries ago [54] whereas, cati species is specific to cats. Both the species have elongated cylindrical morphology [52]. In the start of 1980s, another species was observed with somewhat shorter morphology [53, 54]. It inhabited the outermost layers of epidermis [56] and its length was measured to be 55% to that of canis (**Figure1.12)** [57].

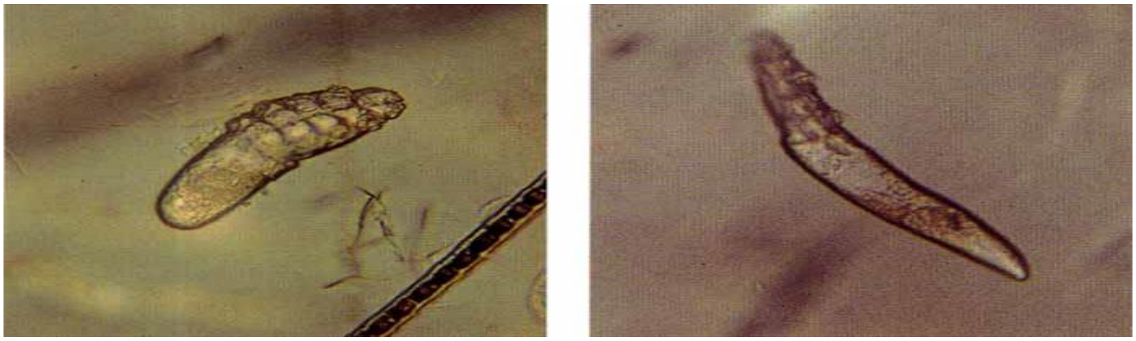


Figure 1.12: A. Short form of Demodex B. Demodex canis

Demodicosis is most prevalent dermal infection, observed at the rate of (5-23%) in nondomestic dogs [58, 59] and (38-58%) in domesticated dogs [60, 61]. The causative agents differ at the start of the infection and at later stages [62].

The infection is categorized as juvenile onset of localized demodicosis (JOLD), in case of having five or less lesions (**Figure 1.14**) [63]. Instead, if the lesions are more than five, especially on feet and legs, before the age of one and a half year, it is categorized as juvenile onset of generalized demodicosis (JOGD) **(Figure 1.15)**. If the symptoms appear after the age of four, the infection is categorized as adult onset of generalized demodicosis **(Figure 1.13)** [65]. The categorization is important for management of the disease [64].

Canine dermatitis is induced when the arthropod spreads in the follicular regions and sebaceous glands [52], which in normal (as in asymptomatic) conditions reside in the skin of dogs [65]. This seems to be caused by poor immune response due to some genetic variants that result in an increase in number of mites than normal [60]. Genetic risk factors appear to be the breed and condition of immune system of the animal [61]. Juvenile onset of localized demodicosis is less severe clinically but may also be followed by bacterial infection so antibiotic may also be needed (**Figure 1.14**) [66]. Generalized demodicosis is more severe and even deadly. In some cases severe pruritus is also observed, hence longtime medication is needed **(Figure 1.15)** [67].

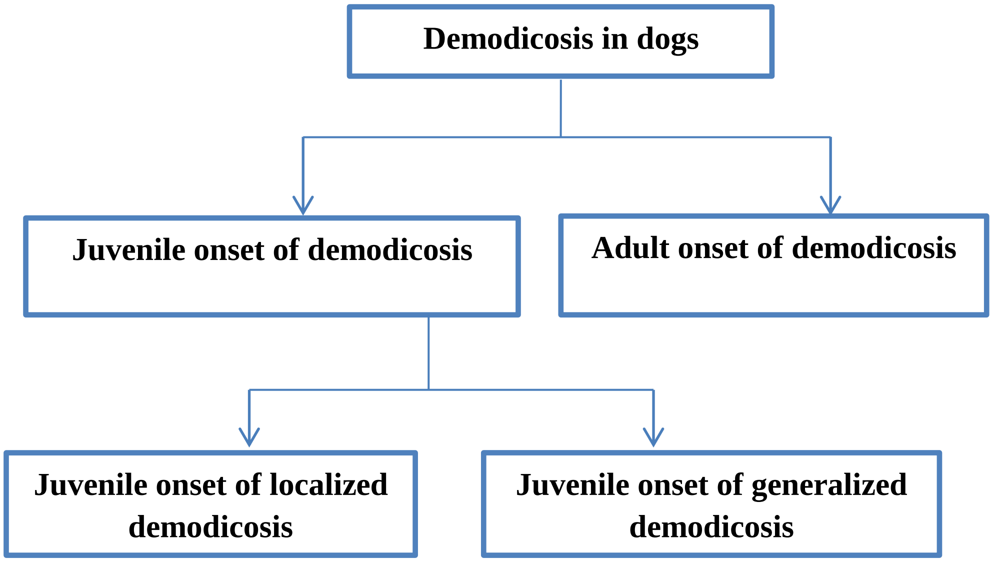

Demodicosis in dogs
Juvenile onset of demodicosis
Adult onset of demodicosis
Juvenile onset of localized demodicosis
Juvenile onset of generalized demodicosis

Figure 1.13: Classification of canine demodicosis

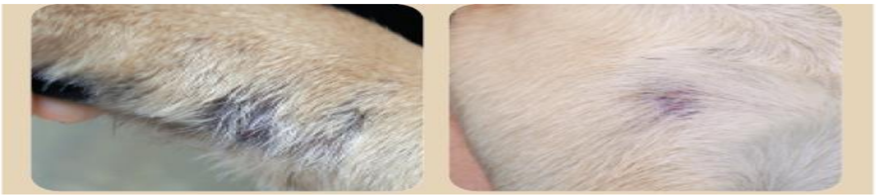


Figure 1.14: JOLD. Symptoms include hair loss and redness of skin

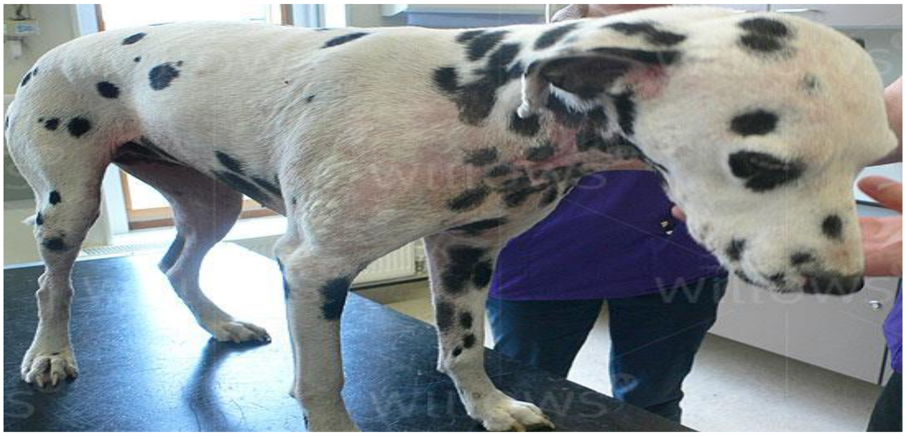


Figure 1.15: JOGD. More invasive than localized type. Symptoms include inflammation, alopecia, and erythema

Factors that make animal susceptible to demodicosis are genetic modifications in cutaneous cellular biochemistry, disorders of immune system, endocrinal status, its age, length of hair coat, stage of reproductive cycle (if it is a female), parturition, internal parasites, and enfeebling diseases [64, 66, 68, 69]. Genetic mutations, as well as acquired inhibition of immune system [64, 70] and consequent cell mediated immune responses can result in abnormal proliferation of demodex mites [71]. Inhibition of immune response of T lymphocytes is due to the presence of α-β globulin found in the sera of infected dogs [72]. On the contrary, some suggest that mites may also stimulate the local inhibition of immune responses [73]. Increased numbers of mites stimulate a humoral factor that suppresses immune system and hence permits the multiplication of mites [74]. The mites also stimulate apoptosis in the infected dogs [75]. They have adapted mechanisms to either stimulate or inhibit apoptosis in host cells and so regulate the immune response [76]. Presence of abnormally lower number of $CD4^+$ T cells, in dogs infected with generalized demodicosis as compared to those with localized demodicosis, may be due to their down regulation by the mites [77].

1.3 Problem Statement

Genetic basis of demodicosis are not known. Hence, the present study aims to make efforts in this direction by investigating the candidate risk factors associated with demodicosis via the GWAS.

1.4 Objectives

The objectives of this study are as follow.

- To examine the single nucleotide polymorphic (SNPs) loci associated with demodicosis in dogs through genome wide association studies.
- To conduct functional analysis of associated SNPs through their annotation.
- To identify biological pathways associated with SNPs to understand their physiological roles.

Chapter 2

Literature review

2. Literature Review

In this chapter the literature search regarding demodicosis is narrated briefly. The sections include a generalized overview of the work conducted with the disease. It is divided into four subsections. The first subsection describes immunological basis of demodicosis. Second includes work cited in the diagnosis and treatment of demodicosis. Third section includes overview of GWAS within dogs. The last section concludes how dogs are suitable model organisms to help identifying risk factors in humans also.

A new species of Demodex was found to be existed in England, Belgium and China. Due to the existence of this species in three different continents of the world, it was considered a common inhabitant of dog's skin. The length of this species was 50 % shorter as compared to the length of female members of Demodex canis [54].

Three breeds were reported to be at highest risk of JOGD namely, American Staffordshire terriers (odds ratio 35.6), Stafford shire bull terriers (odds ratio 17.1) and Chinese shar-pie (odds ratio 7.2) [64]. Induced immunosuppression of dogs was reported to cause development of generalized demodicosis [78, 79]. A dog lost his digit due to podedemodicosis. Skin biopsies confirmed the existence of demodex mites and infestation of demodicosis [80].

2.1 Immunological basis of canine demodicosis

A mechanism of immune response against demodicosis was described as follow. Chitinous skin of mites is recognized by keratinocytes via toll like receptors (TL2). The recognition induces an innate immune response and hence the mite population is controlled in normal circumstances. It has been experimentally validated that immune response involves cellular as well as humoral immunity. It also incorporates the function of CD28 molecules which are co-stimulatory in nature. In case of JOGD, genetic basis are known to occur, but exact abnormality of genome is still unknown. When a dog is suffering from inherited immunological suppression, it is unable to control the mite population. In such a condition, low concentrations of IL2 (interleukin 2) and conversely high concentrations of interleukin 10 (IL10), along with growth factor-β are observed. All these indicate exhaustion of T-lymphocytes. After acarcidal treatment, mites' production starts reducing by killing. The load over T lymphocytes is reduced and hence, the body again starts fighting against the remaining mites. Healing process is also observed as the signs of clinical cure [81].

Strong association of dog leukocyte antigen class II molecules with JOGD was proven through studying the expression of microsatellite markers [82].

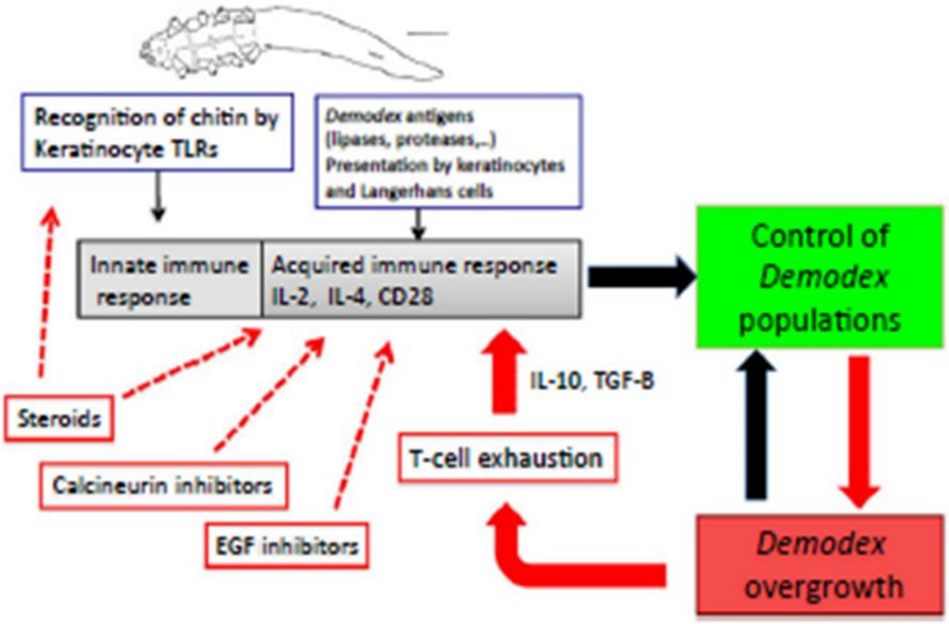

Recognition of chitin by Keratinocyte TLRs
Demodex antigens (lipases, proteases,..) Presentation by keratinocytes and Langerhans cells
Innate immune response
Acquired immune response IL-2, IL-4, CD28
Control of Demodex populations
IL-10, TGF-B
Steroids
Calcineurin inhibitors
EGF inhibitors
T-cell exhaustion
Demodex overgrowth

Figure 2.1: Innate Immunity against demodex mites. Toll like receptors of keratinocytes recognize chitinous skin of the mites and stimulate an innate immune response. Antigen Presenting Cells (APC) of the host recognize foreign lipases, proteases and other antigens to B cells and hence B cells start acquired immune response to kill the antigens.

Considerable decrease in CD4$^+$ T cells to that of CD8$^+$ T cells was observed in dogs infected with generalized demodicosis in comparison to those suffering from localized demodicosis as well as in healthier ones [77].

Significant increase in acute phase C reactive protein was reported in dogs suffering from generalized demodicosis as compared to healthier ones. It was proposed that overpopulated Demodex mites might have the ability to evoke inflammatory immune response in affected dogs [83].

The concentration of IL10 was higher in dogs suffering from recurring infestation of demodicosis as compared to those suffering from the disease for the first time and also the healthier ones [84].

Earlier onset of apoptosis in dogs' peripheral blood leukocytes was proposed to cause immunosuppression in dogs affected with JOGD. This apoptosis irregularity was described to permit the overgrowth of mites in affected dogs [76].

The presence of IgG (immunoglobulin G) was observed in dogs affected with JOGD indicating towards the presence of humoral immune response in affected dogs [85].

An increased $CD8^+$ T lymphocyte count and consequent decline in $Cd4^+$ to $CD8^+$ ratio was suggested to be the potential cause of immunosuppression in dogs suffering from generalized demodicosis [86].

2.2 Diagnosis and treatment of canine demodicosis

It was concluded that definitive doses of Ivermectin (300µg/kg) moxidectin (400µg/kg) and mibemycin (2mg/kg) to be administered orally and Amitraz to be applied on biweekly basis. These doses were suggested to treat generalized form of canine demodicosis. It was further suggested not to breed the dogs infected with generalized demodicosis, because the disease has genetic basis [87].

Ivermectin was proposed to be a better treatment in cases where dogs suffering from JOGD were resisted to Armitaz rinses [88].

Presence of a single mite in dermal skin scraping test or impression tape test was described to be enough to diagnose the infestation of demodicosis [89].

It was demonstrated in a comparative study that adhesive (impression) tape test (ITT) and hair plucking (trichograms) were 75% and 73% sensitive to detect the presence of demodex mites respectively when compared with 100% sensitivity of skin scraping. It was further suggested to use any of the above mentioned two tests because, they were less invasive. The dermal skin scraping test (DSTT) was proposed to be gold standard. The DSST must be used to confirm the presence or absence of demodex mites if ITT and trichograms show negative results [90].

Acetate tape test (ITT) was demonstrated to be significantly reliable procedure to diagnose D. canis as well as S. scabiei species with p-values equivalent to 0.0007. The method was again reported to be less invasive as compared to skin scrapings test (DSST). In addition, Demodex canis mites in different stages of development were observed through impression tape test (ITT). It was also confirmed that there was strong role of heredity as a risk factor to develop demodicosis [91].

2.3 GWAS in dogs

Five distinct blocks of LD in dogs were reported with a size of 5 Mbp each. All these blocks bear five clusters each on chromosome 1,2,3,34,37. These discoveries were concluded by WGA mapping of 20 dogs from 5 breeds. The extent of LD in dogs was also measured to be 100 times greater as compared to humans. The dogs were also shown to possess lower diversity in 2-4 haplotype blocks. These blocks encompass 80% of the dog's genome. In this way merely 15000 SNPs were sufficient to cover these blocks through GWAS [92].

The gene Striatin possesses an 8bp deletion in 3`UTR on chromosome 17. It was proven by another GWA study. The deletion leads to a reduction in mRNA expression of Striatin due to which cardiac muscle fibers are weakened. The deletion is in dominant mode of inheritance. So the disorder becomes more lethal in homozygous condition [93].

The SNPs on *CFA31* within a segment of SODI gene were significantly associated to the canine degenerative myelopathy [94].

In another attempt of GWAS in 20 dogs, a mutation was identified in noncoding 1Mbp region in a locus termed as white spotted. It was reported that *MITF* gene expression is regulated by different combinations of a set of three variants showing phenotypic diversity [95].

A SNP in close proximity of *LHX3* gene was identified on chromosome 9 of the German Shepherd breed. The SNP was associated to a disorder named as pituitary dwarfism. Two SNPs located up and downstream the gene *SODI,* were significantly associated to the another dog disease Degenerative Myelopathy. Many SNPs on chromosome 12 were identified to be significantly associated to the disorder Mega-esophagus and Pancreatic acrinar apathy was having multi genic associations to be significantly involved in the disease [96].

A region on chromosome 27 of dogs and a frame shift mutation in *ADAMTS20* were described to be significantly associated to the cleft palate fetal defect. A SNP within same gene was also identified to be significantly associated to the cleft palate in humans through family based GWAS [97].

2.4 Significance of dogs as model study organisms for humans

Most of the modern dog breeds emerged from bottleneck of two founder populations. One were the domesticated wolves, other was the selective artificial breeding of the original dog genus. As a result of so close ancestry, dogs possess long segments of linkage disequilibrium (LD) in their genomes as compared to those of humans. Furthermore there is far less diversity in their genomes requiring very less number of single nucleotide variants to map across the genome to identify susceptibility loci associated to any disease or phenotype of interest [98].

In contrast to human genome wide association studies, canine genome wide association studies require very small sample size and very small number of markers to encompass whole of the genome. This also makes the study efficient and less expensive. Once a region or a gene is identified in canine GWAS, it can also be finely mapped in a replication study in humans [99].

One of the disadvantages of selective artificial breeding was that they become more prone to genetic disorders like autoimmune diseases, behavioral diseases and cancers, similar to those found in humans. Humans have taken advantage of this mimicry by studying diseases like epilepsy and narcolepsy whose genes were very first time mapped in dogs and then were mapped in humans in a replication study [100].

In another study 114 variants were successfully identified to be significantly associated with obsessive compulsive disorder (OCD) which could not be identified in a human GWAS study with far greater sample size and far more number of variants genotyped [101].

In the similar manner while studying osteosarcoma in dogs, a locus at position rs1906957 found to be on intron of *GRM4* gene, was identified [99]. The gene is involved in cAMP signaling and inhibition. Pathway analysis of this gene and other significant variants in humans revealed strong connections with osteoblasts cell cycle. So it is suggested that canine discoveries of susceptibility loci could be replicated in humans to understand the biology of the disease [99].

It was also proposed that genome wide association study in case of demodicosis in dogs might explore the genetic causes of canine demodicosis [81].

Chapter 3

Methods

3. Methods

The chapter includes materials and methods employed in GWAS analysis to achieve the significant results.

3.1 Pipeline of GWAS

The pipeline of GWAS analysis is beautifully explained in figure 3.1. Selection of cases and controls is carried out from same population to avoid the spurious associations due to differences in ancestry. Cases related to the phenotype of interest should be carefully selected according to standard diagnostic criteria, so that no false associations are identified. The samples are than genotyped using a suitable genotyping platform. The genotypes are than called and subjected to quality control criteria in order to remove poor quality samples as well as SNPs. The extensive statistical analysis results in identification of those SNPs which are significantly associated to the phenotype of interest. The SNPs or markers are than interpreted in order to give them biological meanings and to understand how they could lead to diseased conditions [102].

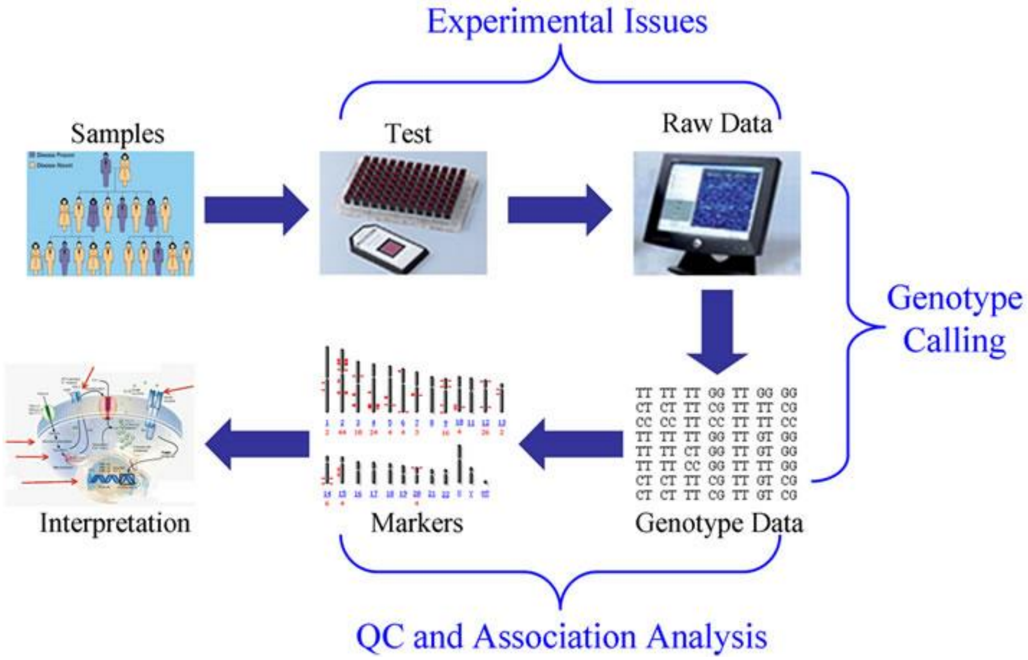


Figure 3.1: Pipeline of steps implemented in GWAS

3.2 Materials and Methods

3.2.1 Subjects of study

In this study we had genotyped data of 188 Stafford Shire Bull Terriers, sampled at Swedish University of Agricultural Sciences, SLU, Uppsala, Sweden. These samples included 94 cases and 94 were controls. The 94 cases were further categorized into 43 localized and 51 generalized. For cases, dogs having signs and symptoms of demodicosis before the age of 18 months were chosen. The phenotype characterization of cases was carefully done according to standard criteria. Those individuals having less than 5 body parts affected were categorized as localized. Those having more than five parts of the body affected were categorized as generalized.

Skin scraping trichograms of Swedish healthy dogs were extracted from three different randomly selected areas of the body identifying no demodex mites on direct microscopy. Gender, age at the onset of the disease, type of disease and case control status was recorded as demographic information. All dogs were genotyped using 170K illumina HD canine SNP array and mapped with Can-Fam 3 Genome Assembly from UCSC Genome browser [103].

3.2.2 Data management & processing using Plink

Data manipulation and trimming was performed using plink which is pioneer in conducting GWAS studies. Plink command --keep list.txt was used to make binary file which was than recoded using --recode and --tab commands with --no web options [106]. The data processing was done using plink software. In doing so, a binary file (.bed) was created by the following commands.

```
plink --file data --allow-no-sex --dog --geno 0.25 --maf 0.05 --mind 0.25--noweb --out data --make-bed [104].
```

Plink efficiently conducts quality control while making the binary file. In this context both the genotypes and individuals having less than 75% call rate are eliminated while the SNPs having minor allele frequency less than 95% are also removed. As a consequence newly generated datasets from this binary ped file already contain a good quality data. Plink generated two respective map and ped files by making a binary file with –keep file.txt command and then with another run of –bfile –recode command [104].

3.2.3 Workflow of GWAS analysis in GenABEL

The steps implemented in GWAS analysis to identify candidate risk factors for juvenile onset of demodicosis in dogs, are expressed through **Figure 3.2.**

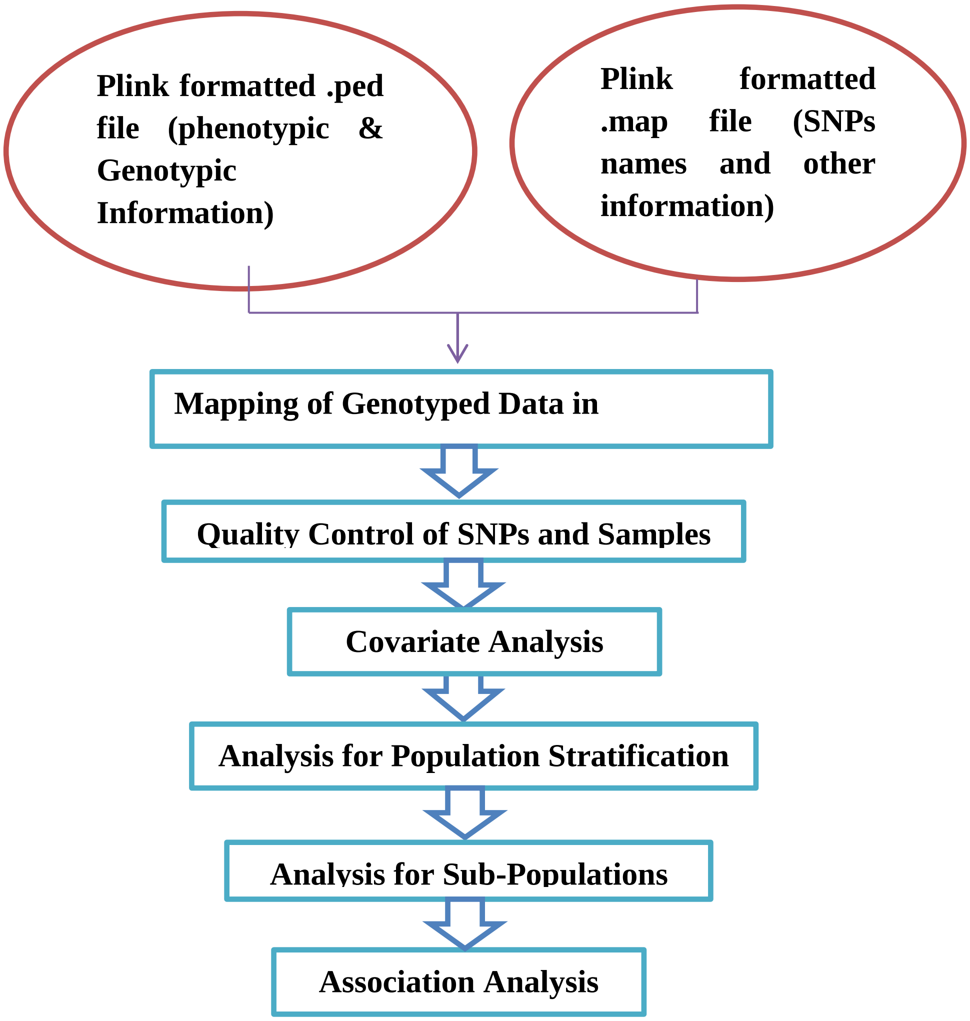

Plink formatted .ped file (phenotypic & Genotypic Information)
Plink formatted .map file (SNPs names and other information)
Mapping of Genotyped Data in
Quality Control of SNPs and Samples
Covariate Analysis
Analysis for Population Stratification
Analysis for Sub-Populations
Association Analysis

Figure 3.2: Workflow of GWAS Analysis

3.2.4 Genome wide association analysis in GenABEL

3.2.4.1 GenABEL data import

GenABEL is an R package, efficiently designed to conduct GWA analysis. Demographic file of the data contains information about individual IDs (identifiers), family IDs, gender, affection status and type. Individuals with male gender were coded as 1 and females as 0 to make it compatible to GenABEL. In a similar manner cases were coded as 1 and controls were coded as 0. All phenotypic information was saved in .dat formatted file. A short overview of .dat file is presented in table 3.1[105].

Table 3.1 GenABEL formatted .dat file

Identifiers	Gender	Affection status (demodicosis)	Type
CFA000208	0	1	2
CFA002888	1	0	0
CFA001551	0	0	0
CFA001663	0	0	0

Gender number 0 is for females 1 is for males. Affection status 0 is for controls and 1 is for cases. Type 1 is for localized and 2 is for generalized

Genotypic information was contained in plink formatted .ped file. .Ped file contains at least six columns containing Individual Ids, Family Ids, Paternal Ids, Maternal Ids, gender and Phenotype of Interest. In contrast to GenABEL, 1 stands for controls while 2 stands for cases in sixth column. All the columns next to phenotype column are the genotypes for SNPs for which all the samples have been genotyped. Therefore, for 188 subjects in the study, there are 188 rows and 6+105769 columns. First six columns contain demographic information while all the rest contain genotypic information for each subject row wise [104]. A short tabular view of plink formatted ped file is presented in **table 3.2.**

Table 3.2 Plink formatted .ped file

IID	FID	PID	MID	Gender	Affstat	SNP1	SNP2	SNP3
CFA000050	CFA000050	0	0	0	2	A A	A G	C C
CFA000207	CFA000207	0	0	0	2	A A	A G	C C

Information of family, paternal, maternal identifiers and gender is compulsory to provide in .ped file. If not available the columns can be filled in as above from 2-5.

The information regarding the markers was stored in plink formatted .map file which must be changed to make it accessible to GenABEL. Plink formatted map file contains four columns each having information about chromosome, markers names, genomic distance in Morgan and base pair position of each marker (SNP) on respective chromosome respectively [104]. For GenABEL compatibility third column containing genomic distance is deleted and headers for the rest three columns are inserted namely chrom (chromosome), marker (SNP), position (genomic base-pair position) respectively [105]. GenABEL formatted map file is shown in **table 3.3** for same first five markers of the data.

Table 3.3 GenABEL formatted .map file

Chrom	Marker	Position
1	TIGRP2P259	249580
1	BICF2G630707908	273487

Chrom is for chromosome, Marker is for SNP name and Position is for base pair position of the SNP on the respective chromosome

Plink formatted .ped and map file are imported into GenABEL through convert.snp.ped command without “makemap” argument. The output file is a .raw file merging genotypic information from both the above mentioned files. The phenotypic dat file and genotypic raw file are loaded into gwaa.daa.class and the output data file is analyzed for genome wide association study in GenABEL. The data file contains row-wise information of SNPs in 14 columns in a sequence as follows, marker, chromosome, position, strand (u for unknown), allele coding A1 or A2 no. of observed genotype, call rate, allele frequency, genotypic distribution (P11 for homozygous effallele, P12 for heterozygous, P22 for homozygous reffallele), Pvalue for the exact test for HWE, Fmax (estimate for deviation from HWE), LRT (P-value for HWE) test are listed [105].

3.2.4.2 Quality control

Preliminary quality control was done to remove the noisy data which otherwise could mislead the analysis by showing spurious associations. GenABEL takes data object of gwaa.data.class and uses the function “check.marker” [105] to extract good quality SNPs and individuals and remove the ones having poor quality. The function check.marker screens the SNPs and individuals on the basis of following parameter as shown in Table 3.4.

Table 3.4: GenABEL Quality Control Parameters

Parameter (function in GenABEL)	Description
Call	threshold call rate for SNPs
perid.call	threshold call rate for individuals
het.fdr	false discovery rate for extraordinary high individual heterozygosity
Ibs	threshold for identity by state to be included
Ibsmark	markers used to estimate identity by state
ibs.exclude	to investigate whether both samples with IBS > ibs be excluded or the one with lower call rate
Maf	minor allele frequency threshold. By default it is 5/2*nids
p.level	Pvlue threshold for Hardy Weinberg equilibrium

Fdrate	FDR threshold for hardy Weinberg equilibrium
odds	odds threshold remove markers that are sex-linked
Hweidsubset	a subset of individuals to be investigated for HWE
Redundant	to check redundancy between chromosome
XXY.call	to check whether the sample is a male or a female by investigating the presence of Y chromosome n proportion to X chromosome
Parameter is the name of parameter used in GenABEL. Description is the explaination of the respective parameter.	

Above mentioned parameters strictly check for poor quality individuals and markers and exclude the ones, which are the result of systematic biases due to batch effects or some other reasons. If not eliminated, such poor quality data are eligible to play the role of confounding. Some of the parameters like SNP call rate, per individual call rate, maf and p-level for HWE are set by the user according to situational requirements.

Quality control is conducted twice. First quality control is conducted on complete dataset to eliminate poor quality SNPs and individuals while second quality control is carried out on the controls only to ensure that controls don't play the role of confounding in the analysis. In doing so, the markers out of HWE are eliminated from the controls through fdrate (false discovery rate) parameter, which sets false discovery rate threshold for identifying markers out of HWE [105].

3.2.4.3 Covariate analysis

Covariate Analysis is important to check if any variable other than affection status to disease is influencing the infestation of the disease in any way. Two tests are conducted to check whether some variables are correlating with each other or not. These are as follow;

3.2.4.3.1 Fisher’s exact test

Fisher’s exact test is a test through which independence of rows and columns in a contingency table is checked. For a 2*2 contingency table p-values are computed using central or non-central hyper-geometric distribution. For larger than 2*2 contingency table and argument (hybrid = TRUE), the asymptotic chi square probabilities are used, provided that no cell has zero count and minimum 5 counts are present in more than 80% of the cells, otherwise exact calculation is used [106-113].

3.2.4.3.2 Pearson product moment correlation test

Pearson product moment correlation test is the method following t-distribution with degrees of freedom 2 less than the length of the object for samples that follows independent normal distribution. Given at least four complete pairs of observations an asymptotic confidence interval is provided [114, 115].

3.2.4.4 Analysis for population structure

Population structure analysis is very important while performing GWAS analysis, because if present, stratification leads to spurious associations. It also results in inflation of the test statistic for associations of SNPs to the disease or phenotype of interest [30]. First step in this regard is the calculation of genomic kinship matrix. This kinship matrix is required to compute genomic differences between samples via autosomal markers. GenABEL has an efficient function "ibs” [105] to calculate identity of state (ibs) between autosomal markers with argument weight = “freq”. IBS values range between -1 to 1. High values of IBS indicate relatedness with 1 being an indicator of twin ship between two samples. During the calculation of ibs, monomorphic SNPs are considered as neutral. The distances between ibs values are stored in a distance matrix and are further used in multidimensional scaling to obtain mds plots [116-124]. These mds plots indicate population stratification, if there are two or more distinct groups within the population. The mds plots are better visualized, when gender and case control status are also mentioned.

3.2.4.5 Subpopulation analysis

Optimum number of sub populations, if present in the data, can be obtained by K means clustering [125-128]. The goodness of a clustering is checked through looking at within cluster sum of squares. Optimum K value is obtained by taking minimum values of within cluster sum of squares (wss).A scree plot of wss Vs. K can be observed for number of subpopulations in the data. The bend in this plot indicates the clusters. After clustering, the individuals of the subpopulations are assigned different vectors accordingly. Coordinates for each individual in the subpopulations are separately stored and population structure is plotted in mds plots. Hence, the mds plots identify any outliers as well as subpopulation in the data, if it exists.

3.2.4.6 Association analysis

There are different approaches to investigate the association of SNPs with phenotype of interest, which is demodicosis in this case. The approaches include genomic control, mixed model approach and structured association approach using principal component analysis, to account for ancestral differences between cases and controls [30].

3.2.4.6.1 Genomic control

GenABEL function “qtscore” [129, 130] is meant for this purpose. This function investigates the association between a SNP and phenotype of interest. With covariates the analysis of the phenotype of interest is carried out through generalized linear model. The residuals of regression are then utilized for test of association of respective SNP and the phenotype of interest. It is carried out through armitage test with 1 degree of freedom. Effects are odds ratios expected in logistic regression model.

Formula used in logistic regression is as follow Y ~ a + b. Genomic inflation factor lambda is estimated as output of the analysis. With the function “estlambda” qqplots are obtained, which are visualization of chi-square distribution of data with expected chi-square values on X-axis while observed chi-square values on Y-axis, the null hypothesis bring no association. Another type of plots called as Manhattan plots are obtained with the function “plot”. The Manhattan plot is a plot showing p-values of all the markers used in the analysis with their chromosome location on X-axis while -ve log of P-values on Y-axis [34].

3.2.4.6.2 Structured association approach

The function implemented in GenABEL for structured association is “egscore” [30]. It efficiently scores association between phenotype of interest and a SNP. However, the important function it performs is the application of principal components to adjust for population stratification. The output is the lambda statistics for one degree of freedom [32,105].

3.2.4.6.3 Mixed model approach

Mixed model approach is used to account for issues of relatedness in the data to minimize the possibility of spurious associations. The GenABEL function “polygenic_hglm” utilizes hierarchical generalized linear model and genomic kinship matrix, obtained by “ibs”, to calculate heritability through extended quasi-likelihood estimates. The estimates are in turn used for estimating restricted maximum likelihood. The polygenic model is finally calculated by taking the covariates as fixed affects [129-138].

The function “mmscore” takes as input the object, returned by “polygenic_hglm” along with gwaa.data object and kinship matrix obtained by “ibs”. The variance of the kinship matirix controls for relatedness. The association is determined by residuals of the hglm model along with the covariates specified in the formula of polygenic function. The output is stored in summary of the “mmscore” object which contains top ten SNPs with lowest p-values their chisquare statistic and other details [139].

3.3 Visualization of results

Manhattan plots and QQ-plots (Quantile Quantile plots) are the two ways which are used to explain GWA results, in addition to summary table showing 10 top most significant SNPs. Manhattan plots are plotted by taking –ve log of p-values of individual SNPs on

Y-axis and chromosomal location of individual SNPs n X-axis. [34] In this way highly significant SNPs can be seen at highest points with respective p-values along Y-axis and

can visually be mapped down to the chromosome they belong to along X-axis. QQ-plot (Quantile-Quantile plot) is another way to visualize the distribution of test statistic as compared to expected test statistic under the null hypothesis of no association [34].

3.4 Biological interpretation of results

GWAS results can be interpreted by pathway analysis, as has been stated in many cases [140-144]. It is better recommended to extend the interpretation of the significant SNPs pathway analysis to investigate the functionality of the genes linked to significant SNPs [145]. For individual SNPs having p-values significant enough, Over Representation Analysis (ORA) is used for pathway analysis [145]. It is also called Functional Enrichment Test. It calculates p-values on the basis of hyper-geometric Test [148]. Pathway analysis is analogus to GO (gene ontology) but is more elaborative and deep [147]. In this context, ANNOAR was used for gene annotation of significant SNPs, which resulted in genes neighboring the significant SNPs. GO terms were used to interpret the gene functions. The SNPs list containing all the ten top most SNPs was obtained through summary of mmscore () object in the GenABEL. The SNP list was passed to ANNOVAR along with reference gene annotation file of canine SNP build Can Fam 3 which was downloaded from UCSC genome browser. Biological interpretation of these genes was done through GO, which is a gene ontology consortium [147].

CHAPTER 4

RESULTS AND DISCUSSIONS

4. Results and Discussions

The chapter discusses the results of GWAS analysis in GenABEL in an attempt to identify risk alleles associated with the juvenile onset of demodicosis. It contains the results starting from quality control, covariate analysis, analysis of population stratification and then the association analysis. It also includes gene annotation and biological interpretation through gene ontology terms.

4.1 Quality control

The analysis started with ~105769 SNPs that were obtained from SNP genotyping platform, genotyped across genomes of all 188 samples including cases and controls. Quality control was carried out twice to avoid any systematic biases and confounding due to batch effects, manual artifacts and failure in genotyping the samples properly. The parameters thresholds were set to be 95% call rate for SNPs and individuals and P = 10e-8 for both minor allele frequency and HWE.

GenABEL performs quality control in two or three (if required) iterations every time resetting the thresholds for those SNPs and individuals who have passed the previous run. In a similar fashion, first quality control was performed on complete dataset containing both case and control groups. In the first run of quality control analysis for whole dataset, 3020 SNPs were removed because they were unable to pass 95% call rate. They were not present in at least 95% of the individuals. Three SNPs were removed in the first run because, they were not successful in reaching the thresholds of < 0.05 for minor allele frequency. They were having very less minor allele frequency count. Total 1430 SNPs were eliminated in the first run of first quality control, because they were out of HWE (p=10e-8). Hardy Weinberg Equilibrium is an indicator of mishandling the samples while sample preparation for genotyping i.e. manual artifacts. Therefore, poor quality SNPs are removed through quality control filters. No individuals were removed that could not possess 95% of the genotyped SNPs implying all individuals passed the filter. After first quality control total of 101490 (95%) of the SNPS and 188 (100%) individuals passed all the first quality control criteria. Mean identity by state (IBS) was calculated to be 0.7143551 (S. e. 0.01591812) and IBS threshold was passed by all the samples.

Table 4.1: Quality Control

	Quality Control1	Quality Control 2
SNP call rate (95%)	**3020**	**_**
Individual call-rate (95%)	**_**	**_**
MAF 0.05	**3**	**2**
HWE (p =1e-8)	**1430**	**1182**
Markers Passed	**101490**	**100306/105769**
Individuals passed	**188**	**188/188**

Second quality control was performed on control group only, in order to remove the SNPs in controls that deviate from HWE. This filtering step is just applied on controls as we assume that in cases deviation from HWE could be the actual signal. In the second quality control, 2 SNPS failed minor allele frequency threshold while 1182 could not reach significance level for HWE and thus were removed. Collectively 100306 (98.83338%) SNPs and all 188 individuals passed all the criteria for quality control 2. Mean IBS was 0.7126142 (s.e. 0.01548113) indicating that both case and control groups were selected from the same group bearing high degree of relatedness. Quality control thresholds and values are summarized in **table 4.1.**

4.2 Covariate analysis

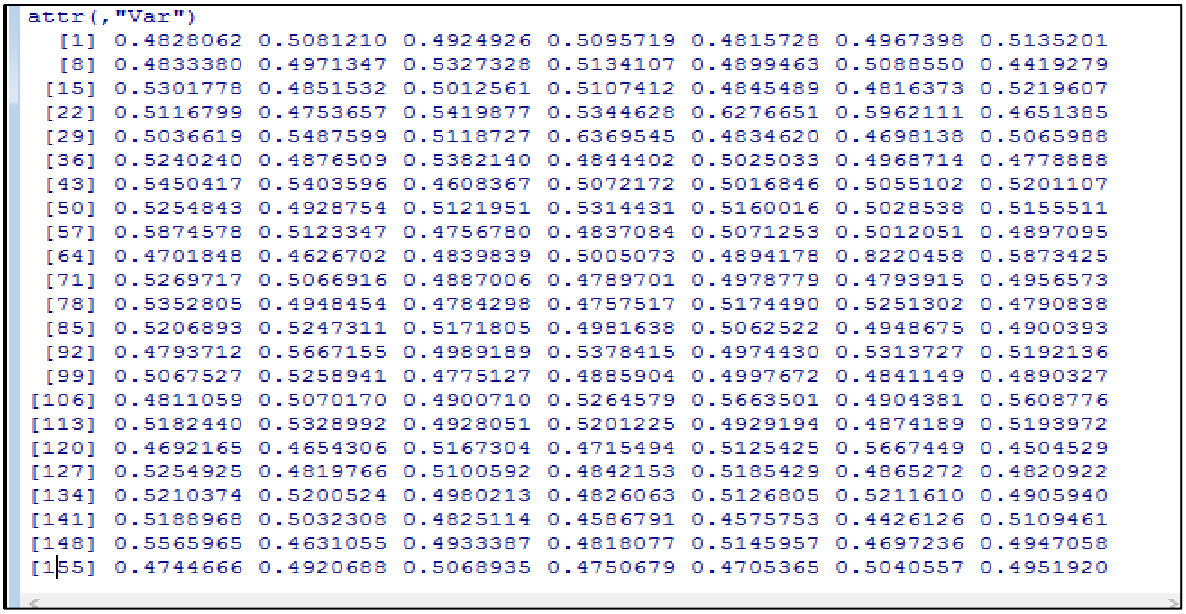
Two different tests were performed to check the status of different covariates present in the phenotype data. Fisher’s exact test was performed to check the status of covariates in the data. While Pearson Product Moment correlation test was performed to observe direct linear relationship of the covariate with affection status that was demodicosis. Results of the covariate analysis are summarized in **table 4.2**. P-values for both tests were > 0.05 and did not show significant results for gender as covariate. It means that gender was not playing the role of covariate in the dataset. Therefore, it was not considered as a covariate in association analysis. However P-values, of both Fisher’s Exact Test and Pearson’s

Product Moment correlation Test were highly significant (P=2.2e-16< 0.05) for type to reject the null hypothesis. Hence, the type was very strongly correlating with the affection status as the value of test statistic for correlation test (0.976) also indicates.

Table 4.2: Covariate Analysis		
Variables to Affection status	**Fisher's exact test**	**Pearson Product Moment Correlation test**
Gender	**0.8824**	**0.7687**
Type(localized, generalized)	**2.2e-16**	**2.2e-16**

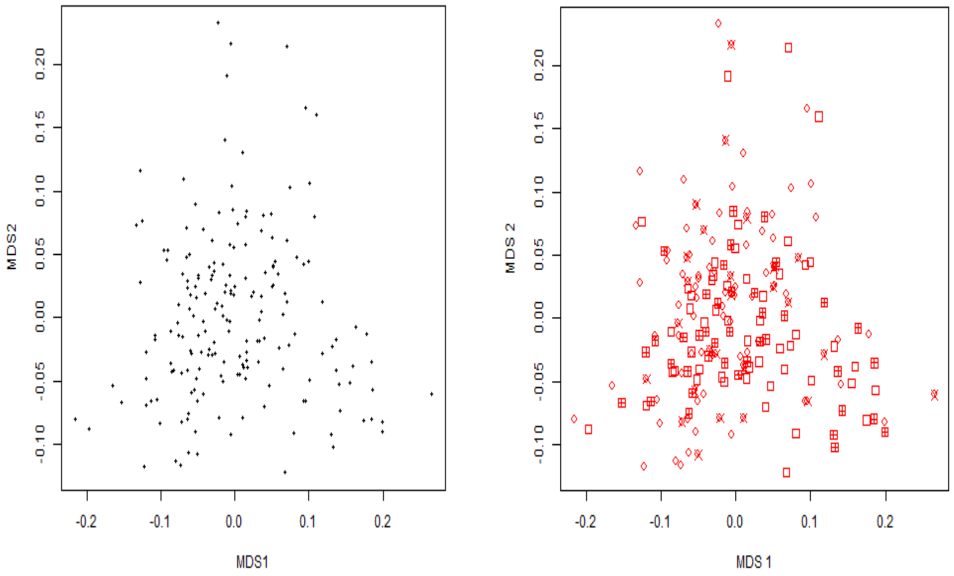
4.3 Population Structure Analysis

Genomic kinship matrix and MDS (multidimensional scaling) plots were obtained taking into account 99724 autosomal markers for investigation of population structure in the data. Results of kinship coefficient are summarized in figure 4.1. Genomic kinship matrix is obtained by calculating similarity of genotypes within any two pair of the population and gives kinship coefficients which thus indicate degree of relatedness between the members of a pair. Positive value indicates relatedness while negative values of coefficient of kinship are indicator of high genomic differences in ancestry. The dataset used, did not show any negative values of kinship coefficient. It was calculated by measuring identity by state of autosomoal SNPs only, because the sex linked SNPs are highly varied even among the individuals of a family and then among the populations of same ancestry [148, 149, 150]. Hence, they are not considered for calculating kinship matrix. Negative values in the matrix are the signs of different ancestry indicating no identity by state. Absence of any negative values in genomic kinship matrix suggests that there was no population stratification [148, 149 and 150]. MDS plot obtained by computing distances on the basis of autosomal markers also present homogeneous populations of both genders and also case and control groups. Thus it can be concluded that there was no population stratification in the data (Figure 4.2).



```
attr(,"Var")
   [1] 0.4828062 0.5081210 0.4924926 0.5095719 0.4815728 0.4967398 0.5135201
   [8] 0.4833380 0.4971347 0.5327328 0.5134107 0.4899463 0.5088550 0.4419279
  [15] 0.5301778 0.4851532 0.5012561 0.5107412 0.4845489 0.4816373 0.5219607
  [22] 0.5116799 0.4753657 0.5419877 0.5344628 0.6276651 0.5962111 0.4651385
  [29] 0.5036619 0.5487599 0.5118727 0.6369545 0.4834620 0.4698138 0.5065988
  [36] 0.5240240 0.4876509 0.5382140 0.4844402 0.5025033 0.4968714 0.4778888
  [43] 0.5450417 0.5403596 0.4608367 0.5072172 0.5016846 0.5055102 0.5201107
  [50] 0.5254843 0.4928754 0.5121951 0.5314431 0.5160016 0.5028538 0.5155511
  [57] 0.5874578 0.5123347 0.4756780 0.4837084 0.5071253 0.5012051 0.4897095
  [64] 0.4701848 0.4626702 0.4839839 0.5005073 0.4894178 0.8220458 0.5873425
  [71] 0.5269717 0.5066916 0.4887006 0.4789701 0.4978779 0.4793915 0.4956573
  [78] 0.5352805 0.4948454 0.4784298 0.4757517 0.5174490 0.5251302 0.4790838
  [85] 0.5206893 0.5247311 0.5171805 0.4981638 0.5062522 0.4948675 0.4900393
  [92] 0.4793712 0.5667155 0.4989189 0.5378415 0.4974430 0.5313727 0.5192136
  [99] 0.5067527 0.5258941 0.4775127 0.4885904 0.4997672 0.4841149 0.4890327
 [106] 0.4811059 0.5070170 0.4900710 0.5264579 0.5663501 0.4904381 0.5608776
 [113] 0.5182440 0.5328992 0.4928051 0.5201225 0.4929194 0.4874189 0.5193972
 [120] 0.4692165 0.4654306 0.5167304 0.4715494 0.5125425 0.5667449 0.4504529
 [127] 0.5254925 0.4819766 0.5100592 0.4842153 0.5185429 0.4865272 0.4820922
 [134] 0.5210374 0.5200524 0.4980213 0.4826063 0.5126805 0.5211610 0.4905940
 [141] 0.5188968 0.5032308 0.4825114 0.4586791 0.4575753 0.4426126 0.5109461
 [148] 0.5565965 0.4631055 0.4933387 0.4818077 0.5145957 0.4697236 0.4947058
 [155] 0.4744666 0.4920688 0.5068935 0.4750679 0.4705365 0.5040557 0.4951920
```

Figure 4.1: Genomic kinship Matrix



A **B**

Figure 4.2: MDS plot showing no stratification in the data

A: homogeneous population B: case (crossed) and controls male and female are constituting a homogeneous population. No sign of stratification is evident from the plots

4.4 Subpopulation Analysis

Subpopulation analysis was performed through plot for MDS and clustering vs. within cluster sum of squares with optimum no of clusters indicating subpopulations**.** The curvature in the plot showed subpopulations at the level of 2 indicating towards the presence of optimally two subpopulations only i.e. of cases and controls. The Multi-dimensional scaling plot obtained in same step indicates that the two subpopulations were homogeneous and no stratification existed in the dataset and all the cases and controls were within same single cluster as far as association analysis with respect to strata in the population is concerned.

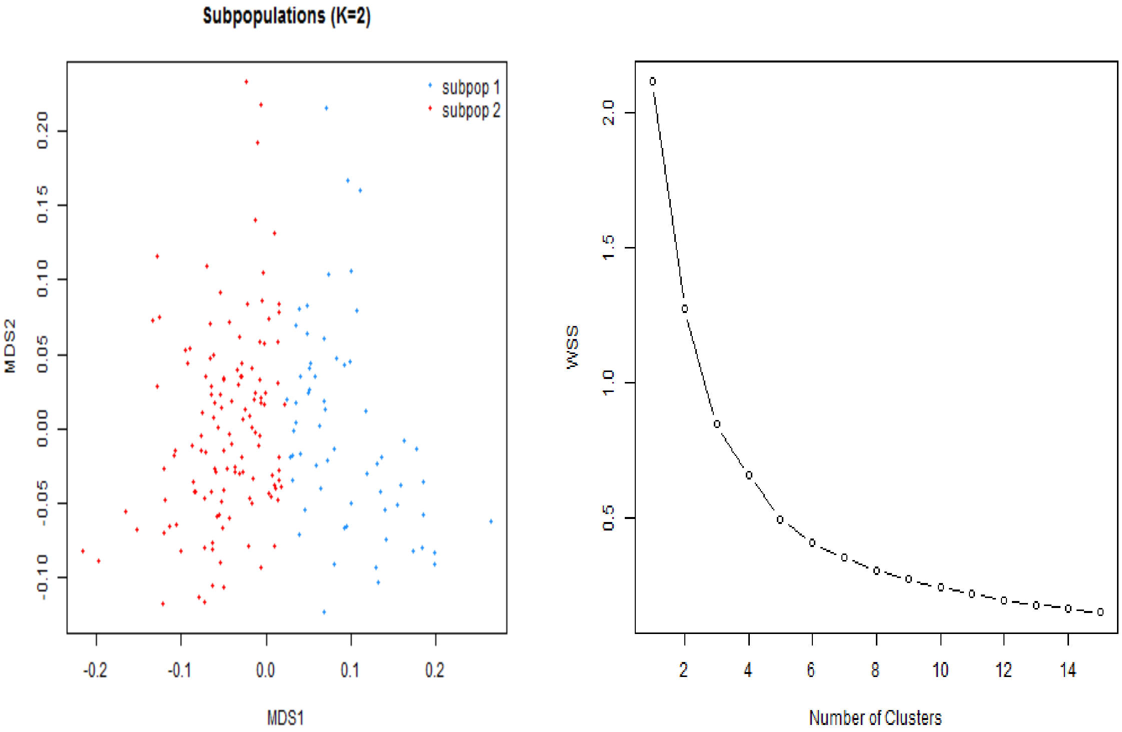


Figure 4.3 A & B: MDS & Scree plot to show subpopulations

4.5 Association Analysis

There were no strata in the data, therefore principal component analysis and analysis of any of the model of association for strata, taking into account the populations, was not done. Simple association and genomic control (qtscore function in GenABEL) was performed to investigate for any association between phenotype and genotype. Association analysis was performed by taking into account type (localized or generalized) as covariate using the function of qtscore for genomic control (GC).The inflation factor lambda was very high up to 1.43, as is also evident by extent of noise in association signals in Manhattan plot (Figure 4.4). However QQ-plot with pc1df (corrected p-value for genomic control) showed much deviation from the line of equality to null in both the cases (with p1df and pc1df). It was indicating cryptic relatedness as there was no apparent stratification in the dataset (Fig 4.5) [102].

Table 4.3 Association Analysis using type as covariate

Model	**Covariate**	**Lambda value**
Simple association	Type	1.420729
Genomic Control	Type	1.420729
Mixed Model	Type	1

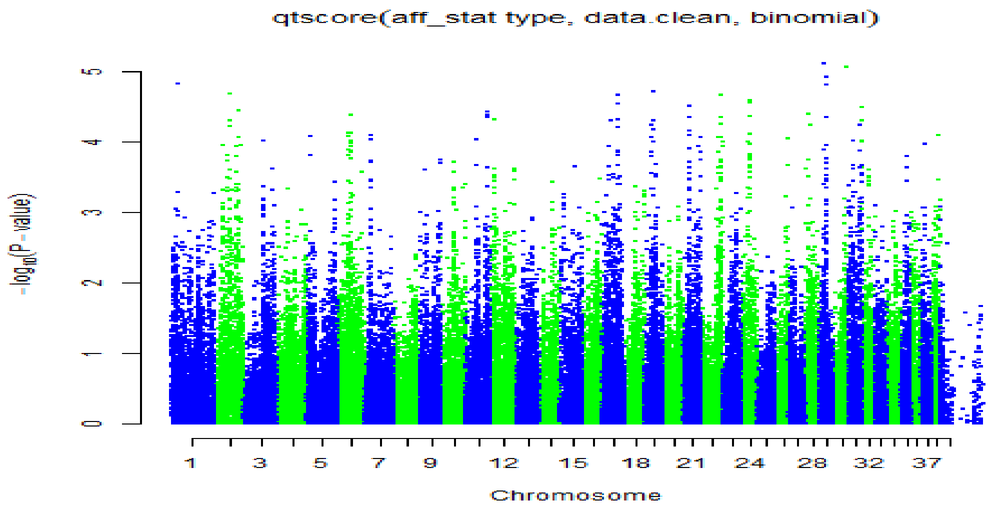


Figure 4.4: Manhattan plots showing noise in the SNPs p-values due to relatedness.

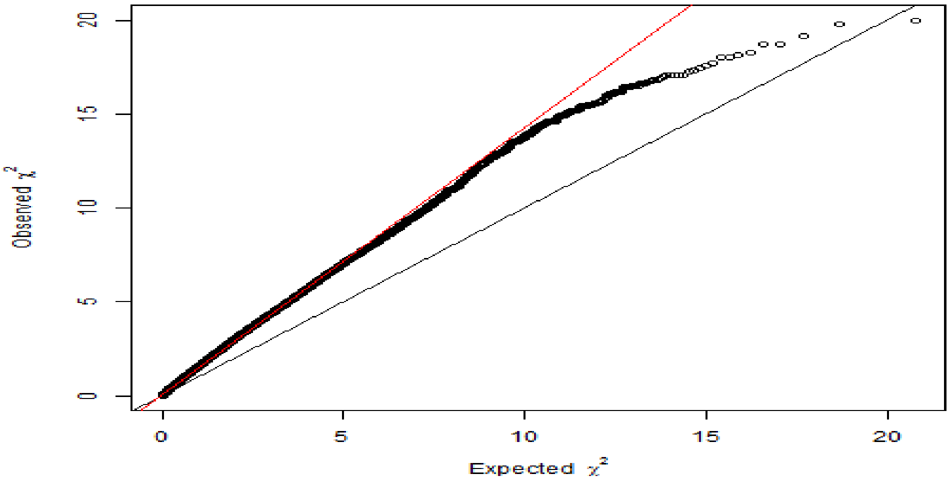


Figure 4.5: QQ-plot showing deflation due to relatedness with qtscore

To minimize this inflation due to cryptic relatedness, mixed model approach was utilized [148, 149 and 150]. In this context GenABEL function "mm-score" was applied taking type as covariate [132-141]. It worked and most significant results of the study were obtained which were supported by all the three criteria used to identify risk alleles in GWAS i.e. Lambda value was equal to 1. In addition both the QQ-plot and Manhattan plot were supporting the SNPs to be true hits (**Figure 4.6 A & B**) [34]. Manhattan plot of the mixed model showed clear signal of all ten top most SNPs on chromosome 28 reaching negative log of p-values up to 5 and even 6, explaining the significance of p-values of top hits, which are summarized in table 4.4. The QQ plot expressed the chi-square distribution of all the SNPs on line of equality, with only the third quantile deviating the null hypothesis of no associations, for significant SNPs, from the line of equality. These SNPs were having chi-square values from 18 to 20, as is also evident for the significant SNPs in **table 4.4**. All top candidate SNPs significantly associated with demodicosis in the study, were located on chromosome 28. The p-value of SNP with highest level of significance was 3.78 e-6 and the p-values of all the rest of the nine significant SNPs was in the range of 10e-5 **(Table 4.4),** showing that none reached Bonferroni corrected threshold (p=0.05/100347=4.98271e-07). However, Bonferroni threshold is very strict and shows conservative approach [102] when dealing with dogs due to very high linkage disequilibrium indicating that the number of independent SNPs is very low in case of dogs [92].

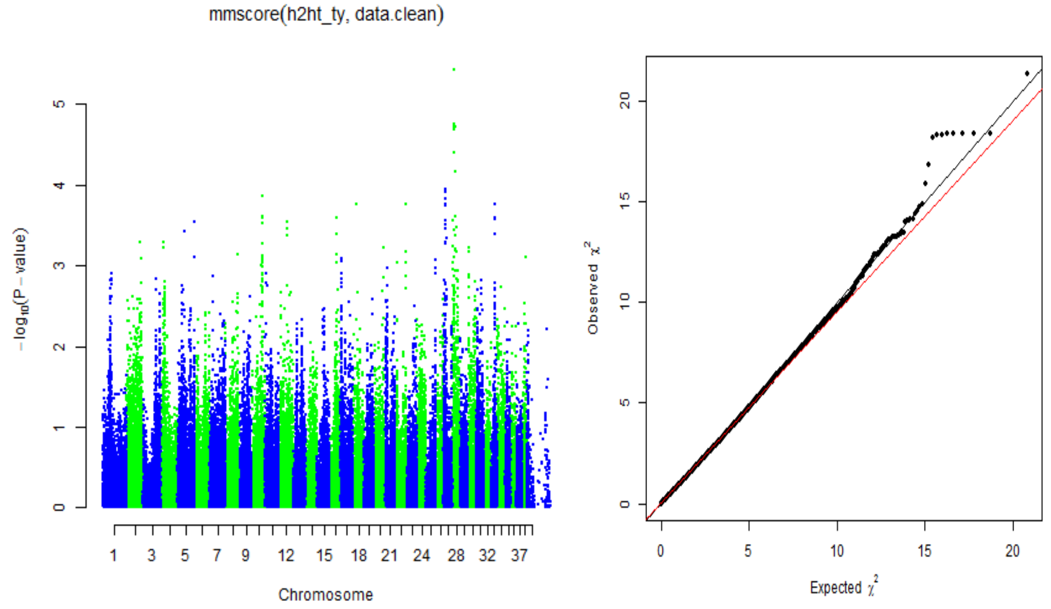
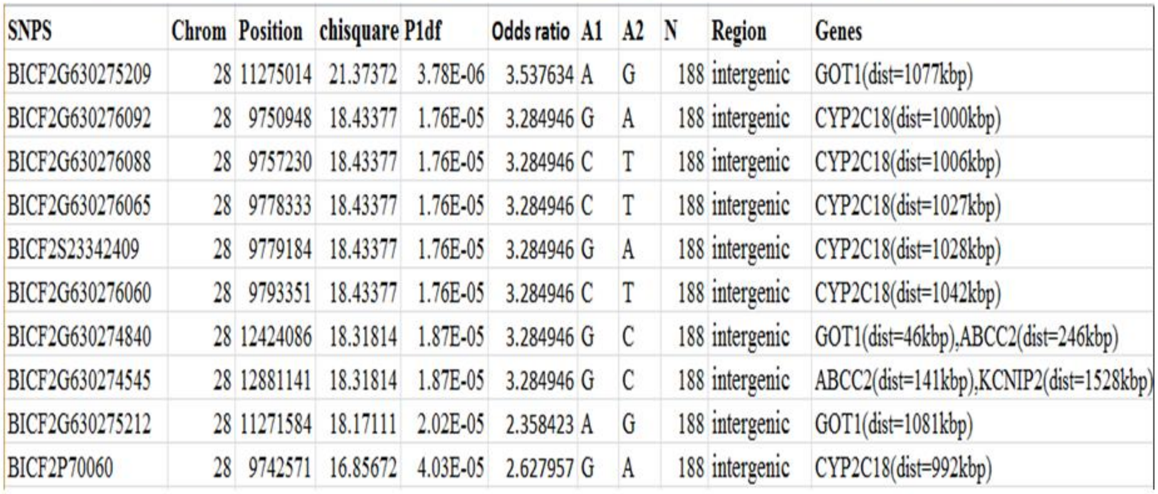


Figure 4.6 A) Manhattan plot showing chromosome 28 to possess significant SNPS with p-values 10e-6 and 10e-5 respectively.

B) QQ-plot showing the deviation of significant SNPs from distribution line which is based on null hypothesis. It clearly indicates how strongly the SNPs are associated to demodicosis (the phenotype of interest) rejecting the null hypothesis of no association.

Table 4.4: Summary of top ten candidate SNPs n (mixed model with type as covariate)



SNPS	Chrom	Position	chisquare	P1df	Odds ratio	A1	A2	N	Region	Genes
BICF2G630275209	28	11275014	21.37372	3.78E-06	3.537634	A	G	188	intergenic	GOT1(dist=1077kbp)
BICF2G630276092	28	9750948	18.43377	1.76E-05	3.284946	G	A	188	intergenic	CYP2C18(dist=1000kbp)
BICF2G630276088	28	9757230	18.43377	1.76E-05	3.284946	C	T	188	intergenic	CYP2C18(dist=1006kbp)
BICF2G630276065	28	9778333	18.43377	1.76E-05	3.284946	C	T	188	intergenic	CYP2C18(dist=1027kbp)
BICF2S23342409	28	9779184	18.43377	1.76E-05	3.284946	G	A	188	intergenic	CYP2C18(dist=1028kbp)
BICF2G630276060	28	9793351	18.43377	1.76E-05	3.284946	C	T	188	intergenic	CYP2C18(dist=1042kbp)
BICF2G630274840	28	12424086	18.31814	1.87E-05	3.284946	G	C	188	intergenic	GOT1(dist=46kbp),ABCC2(dist=246kbp)
BICF2G630274545	28	12881141	18.31814	1.87E-05	3.284946	G	C	188	intergenic	ABCC2(dist=141kbp),KCNIP2(dist=1528kbp)
BICF2G630275212	28	11271584	18.17111	2.02E-05	2.358423	A	G	188	intergenic	GOT1(dist=1081kbp)
BICF2P70060	28	9742571	16.85672	4.03E-05	2.627957	G	A	188	intergenic	CYP2C18(dist=992kbp)

Thus the candidate SNPs and risk loci may be involved in demodicosis in dogs and need further experimental validation, to see the effects of SNPs at risk loci in cases and controls. Although p-values is low for top SNPs but previous studies indicate that low p-values in GWAS may still be significantly associated with risk loci [36, 151, 152, and 155, 156].

All ten top candidate significant SNPs were present in all 188 members of the study. The odds ratio of first nine significant SNPs was 3.53 while that of 10th significant SNP was 2.627 which also proved that all these significant SNPs were associated to phenotype of interest in the study that is juvenile onset of canine demodicosis. Moreover the distance of nearby genes from the tag SNPs is only in hundred kilo basepair (kb) positions for five of the SNPs indicating strong evidence towards the involvement of regulatory regions of nearby genes, obtained through gene annotation via ANNOVAR.

4.6 Pathway analysis

The results of gene annotation showed that all ten SNPs were present in the intergenic region. In such a situation when associated SNPs lie in intergenic regions, it is recommended to study neighboring genes because some non-coding elements such as activators or suppressors may alter the functioning of these genes [157, 158]. Therefore, the SNPs explain the association of these genes to the phenotype of interest.

The neighboring genes include *GOT1*(Glutamic-Oxaloacetic Transaminase 1), *ABCC2* (ATP Binding Cassette Subfamily C Member 2), *KCNIP2 (*Potassium Voltage-Gated Channel Interacting Protein 2*)* and *CYP2C18* (Cytochrome P450 Family 2 Subfamily C Member 18). Five of the SNPs lie in same intergenic region neighboring *CYP2C18* **(Table 4.4)**. *GOT1* is the neighboring gene of top most significant SNP i.e. BICF2G630275209. All four genes were given to gene ontology (GO) consortium for over representation analysis (ORA) and without Bonferroni correction, following results were obtained [151].

GOT1 (Glutamic-Oxaloacetic Transaminase 1) was annotated with a p-value less than 0.05, without Bonferroni correction. It was annotated at the levels; cellular component, biological process and molecular level, with; mitochondrion, cellular amino acid biosynthesis, trans-aminase and ATPase activity respectively. All four genes were involved in catalytic and transporter activity. It is also shown by the pie chart in **figure 4.5**. *KCNIP2 (*Potassium Voltage-Gated Channel Interacting Protein 2*)* was involved in calcium mediated signaling and response to toxins. *ABCC2* (ATP Binding Cassette Subfamily C Member 2) was involved in extracellular transport and *CYP2C18* (Cytochrome P450 Family 2 Subfamily C Member 18) was involved in fatty acid metabolic process [151].

;

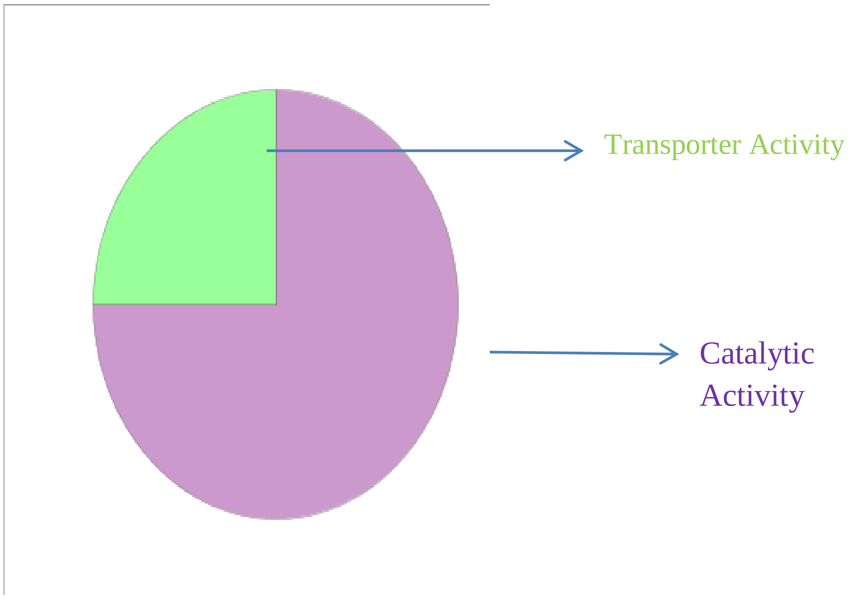


Figure 4.7 Pie Chart obtained by GO terms with a list containing GO Ids of all four genes

4.7 Discussion

The results of the study show that minor alleles for top hit SNPs in the cases may cause suppression in catalytic activity leading to slowing down of immune response. As a result the dogs fail to control the population of mites and suffer from juvenile onset of demodicosis. Presence of tag SNPs flanking the region, containing gene *CYP2C*18 (base pair position from 9750949 and 9793351) may indicate the involvement of a large intergenic segment of chromosome 28 to mediate the disease. All top ten hit SNPs obtained as a result of association analysis, through applying mixed model with type as a covariate, are lying in the inter-genic region of four genes which are; *CYP2C18, GOT1, ABBC2 and KCNIP2*.

The gene *CYP2C18* is in close vicinity of six significant SNPs i.e. top most SNP number 2, 3, 4, 5, 6 and 10 namely BICF2G630276092, BICF2G630276088, BICF2G630276065, BICF2S23342409, BICF2G630276060 and BICF2P70060 respectively from nucleotide position 992144bp till 1042924bp. The second gene on the chromosome 28, neighboring top SNPs is *GOT1* which is at a distance of 1077644bp from the top most significant SNP of our GWAS analysis. This top most SNP is named as BICF2G630275209. Same gene *GOT1* is in close vicinity of another significant SNP BICF2G630275212. The seventh and eighth top most SNPs of our results i.e. BICF2G630274840, BICF2G630274545 are located in the neighborhood of another gene *ABCC2* with a distance of 246799bp and 14101bp respectively. The fourth gene neighboring the 8th significant SNP (BICF2G630274545) is *KCNIP2* which is located 528385bp away from this SNP. The order of these genes on chromosome 28 is *CYP2C18, GOT1, ABCC2* and *KCNIP2*. Results of literature search to determine previously stated biological role of the genes that are neighboring the top candidate SNPs, are summarized in table 4.5. The summary includes; the gene, number of top SNPs located in vicinity of that gene, diseases associated to the gene as reported by gnenecard information and phenotype that are linked to these diseases. In this way the genes were better interpreted to relate to the juvenile onset of canine demodicosis.

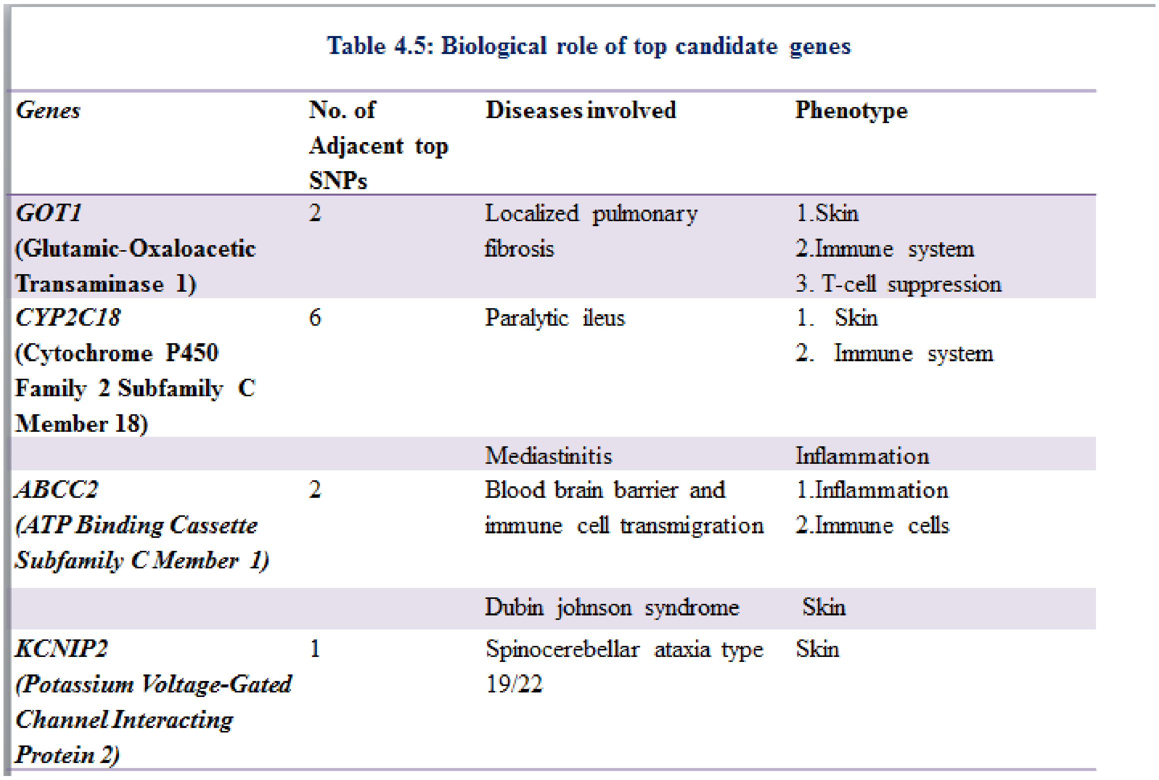


Table 4.5: Biological role of top candidate genes

Genes	No. of Adjacent top SNPs	Diseases involved	Phenotype
GOT1 **(Glutamic-Oxaloacetic Transaminase 1)**	2	Localized pulmonary fibrosis	1.Skin 2.Immune system 3. T-cell suppression
CYP2C18 **(Cytochrome P450 Family 2 Subfamily C Member 18)**	6	Paralytic ileus	1. Skin 2. Immune system
		Mediastinitis	Inflammation
ABCC2 (ATP Binding Cassette Subfamily C Member 1)	2	Blood brain barrier and immune cell transmigration	1.Inflammation 2.Immune cells
		Dubin johnson syndrome	Skin
KCNIP2 (Potassium Voltage-Gated Channel Interacting Protein 2)	1	Spinocerebellar ataxia type 19/22	Skin

CYP2C18 is a member of cytochrome P450 family 2. Its subfamily is C and it is 18th member of this subfamily. It is expressed into an enzyme of monoxygenic nature. The enzyme is occupied on endoplasmic reticulum and is involved in anabolism of lipids and metabolism of drugs. Diseases associated to *CYP2C18* are paralytic ileus and mediastinitis. Paralytic ileus is related to interstitial cystitis whose related phenotypes are skin and immune system. Both the skin and immune system are also related to demodicosis. Mediastinitis is another inflammatory disease which is related to *CYP2C18* gene [159] just like demodicosis. Mutations in this genes have also been reported to be involved in poor metabolism of drugs like warfarin [102]

Hence, *CYP2C18* might be a strong candidate gene responsible for etiology of demodicosis. Furthermore this gene is located in close vicinity of six of the ten top most SNPs. Additionally these 6 SNPs (map position 84087–84262) are very closely located in the map file within a range of 10 tag SNPs indicating that more SNPs with a strong linkage disequilibrium to these tag SNPs with the same haplotype block might be involved in causing the Juvenile onset of demodicosis disease. Therefore, it is a better option for future genotype imputation studies so that actual causal loci might be identified. In short *CYP2C18* gene might be a strong candidate gene associated with juvenile onset of canine demodicosis disease [159].

GOT1 is the gene neighboring the top most SNP of our results. It translates into an enzyme glutamic oxaloacetic transaminase present in soluble form in cytosol. It is an active participant of amino acid metabolism and urea and TCA cycle. One of its related diseases is localized pulmonary fibrosis, which is related to pulmonary fibrosis. Its related phenotypes include integument and immune system. Furthermore, biological processes related to this disease include, suppression of T-cells and stimulation of collagen biosynthesis process which are again related to demodicosis. In addition, the **figure 4.8**, which is a screenshot of gene card results for *GOT1* gene, indicated that the protein is expressed in almost all types of immune cells including T-lymphocytes, CD4$^+$ T Cells and CD 8$^+$ T cells which are found associated with juvenile onset of demodicosis disease [160].

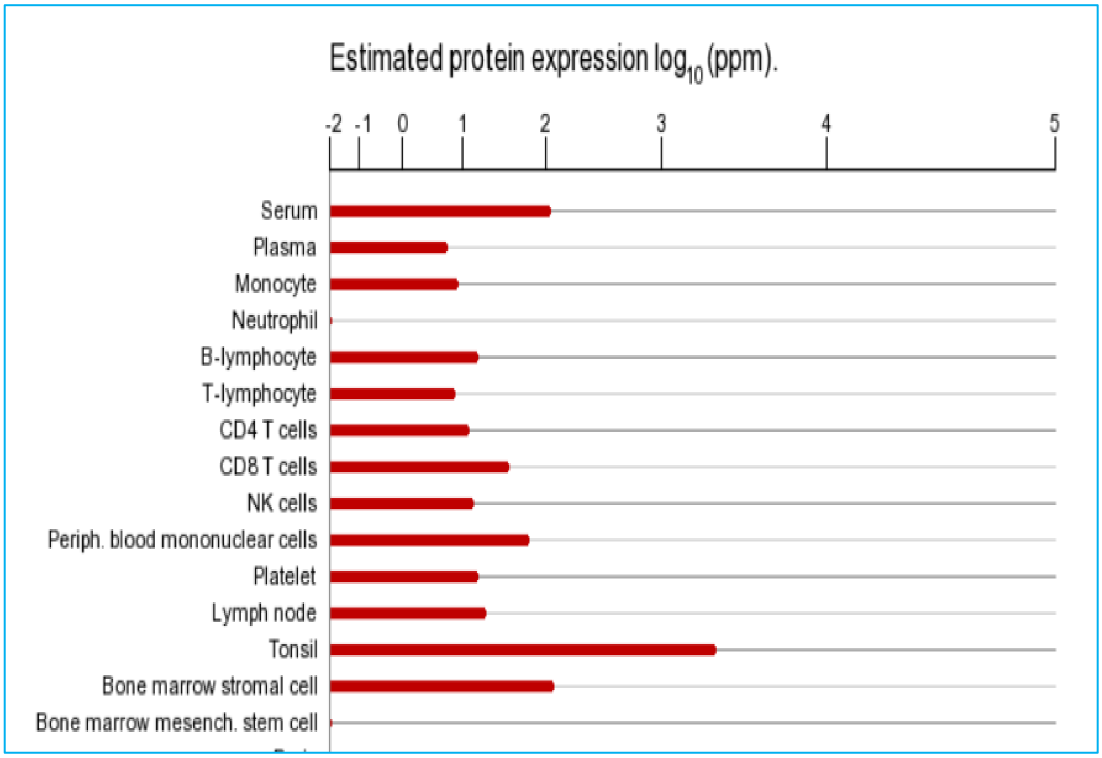


Figure 4.8: Genecard for GOT1: Expression of GOT1 in Immune system

Hence there might be role of this gene associated with demodicosis [160].

ABCC2 is also a protein coding gene. It is translated into a protein named as ATP binding cassette with a subfamily C and member 2. It acts as a transporter across cell membranes. It is expressed in upper portion of liver cells canals and plays a role in transport of bile. It is also involved in resistance of drugs. It is important to discuss here the pathway for *ABCC2* gene namely, “blood brain barrier and immune cell transmigration”. The pathway if disturbed, results in inflammation and invasion of immune cells in brain. One of its related diseases is Dubin Johnson syndrome which is related to hepatitis (the inflammation of liver). Integument is the phenotype associated to hepatitis as is of demodicosis. Its biological processes include positive regulation of collagen biosynthesis [161].

KCNIP2 encodes the protein known as Potassium voltage gated channel interacting protein. It is involved in the regulation of neuronal stimulation. It belongs to the family of calcium binding proteins. One of the associated diseases is spinocerebellar ataxia type 19/22 whose associated tissue also includes skin [162].

All these genes are showing signs of relevance to the juvenile onset of demodicosis disease, to a larger or a smaller extent, by either causing an inflammatory disease or by having skin or immune system as a phenotype. It may be hypothesized that *CYP2C18* and *GOT1* might be strongly associated to the infestation of JOGD while the *ABCC2* may also be indirectly involved in the infestation of this disease.

Chapter 5

Conclusions

5. Conclusion

In this study we performed GWAS on genotyped samples of Stafford Shire Bullterriers in an attempt to identify risk factors for Juvenile onset of canine demodicosis disease. Through applying logistic regression based linear mixed model with type as predictor, we were successful in achieving ten significant SNPs with association P-values in the range of 10e-5 for nine SNPs while the p-value for top SNP was 1.78e-6. The QQ Plot and the Manhattan Plot also supported the significance of these results. All the ten significant SNPs are occupying canine chromosome 28. Gene annotation through ANNOVAR revealed the SNPs to occupy an intergenic region in a vicinity of four genes namely *GOT1, CYP2C18, ABCC2 & KCNIP2.* Two of these genes i.e. *GOT1 and CYP2C18* seem to be involved in the infestation of juvenile onset of generalized demodicosis as they are found to have an inflammatory background. In addition, nine of the significant SNPs in our study are neighboring these genes. Therefore, we conclude that the ten significant SNPs table 4.4 are strongly associated to the phenotype of interest i.e. juvenile onset of canine demodicosis in a discovery stage.

5.1 Future perspective

The study can provide basis for replication studies with independent genotyped samples, in an attempt to validate these results. We can further explore the intergenic region identified here to find some regulatory elements to identify if they are controlling the expression of these genes. This can be achieved using the targeted sequencing of the region in large number of individuals. Same strategy can also be applied to human samples to identify the risk factors associated with demodicosis or associated diseases in humans.

Bibliography

1. Bush, William S., and Jason H. Moore. "Genome-wide association studies." PLoS Comput Biol 8.12 (2012): e1002822.
2. Manolio, Teri A., and Francis S. Collins. "Genes, environment, health, and disease: facing up to complexity." Human heredity 63.2 (2007): 63-66.
3. Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nature Reviews Genetics, 11(5), 356-366.
4. van der Sijde, M. R., Ng, A., & Fu, J. (2014). Systems genetics: from GWAS to disease pathways. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842(10), 1903-1909
5. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. and Hunt, S.E., 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), pp.928-933.
6. Almontashiri, N. A., & Hannan, M. (2015). Usefulness of genome-wide association studies to identify novel genetic variants underlying the plasma lipoprotein metabolism as risk factors for CAD. Journal of Taibah University Medical Sciences, 10(3), 266-270.
7. Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S. and Papadakis, K., 2009. Eight blood pressure loci identified by genome-wide association study of 34,433 people of European ancestry. *Nature genetics*, *41*(6), p.666.
8. Li, S., Qian, J., Yang, Y., Zhao, W., Dai, J., Bei, J.X., Foo, J.N., McLaren, P.J., Li, Z., Yang, J. and Shen, F., 2012. GWAS identifies novel susceptibility loci on 6p21. 32 and 21q21. 3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. PLoS Genet, 8(7), p.e1002791.
9. Visscher, P. M. (2008). Sizing up human height variation. Nature genetics, 40(5), 489-490.

10. Billings, L. K., & Florez, J. C. (2010). The genetics of type 2 diabetes: what have we learned from GWAS? Annals of the New York Academy of Sciences, 1212(1), 59-77.
11. Sullivan, P. F. (2010). The psychiatric GWAS consortium: big science comes to psychiatry. Neuron, 68(2), 182-186.
12. Iles, M. M. (2008). What can genome-wide association studies tell us about the genetics of common disease?. PLoS Genet, 4(2), e33.
13. Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., ... & Bracken, M. B. (2005). Complement factor H polymorphism in age-related macular degeneration. Science, 308(5720), 385-389.
14. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences, 106(23), 9362-9367.
15. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. and Parkinson, H., 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, *42*(D1), pp.D1001-D1006.
16. International HapMap Consortium. (2005). A haplotype map of the human genome. Nature, 437(7063), 1299-1320.
17. Altshuler, D., & Daly, M. (2007). Guilt beyond a reasonable doubt. Nature genetics, 39(7), 813-815.
18. Wray, N. R., Goddard, M. E., & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome research, 17(10), 1520-1528.
19. Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon et al. "Initial sequencing and analysis of the human genome." Nature 409, no. 6822 (2001): 860-921.
20. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. Nature genetics, 29(2), 229-232.

21. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. and Liu-Cordero, S.N., 2002. The structure of haplotype blocks in the human genome. Science, *296*(5576), pp.2225-2229.
22. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., ... & Kruglyak, L. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science, 280(5366), 1077-1082.
23. Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. TRENDS in Genetics, 17(9), 502-510.
24. McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science, 304(5670), 581-584.
25. Syvänen, A. C. (2005). Toward genome-wide SNP genotyping. Nature genetics, 37, S5-S10.
26. Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science, 273(5281), 1516-1517.
27. Hunter, D. J., & Kraft, P. (2007). Drinking from the fire hose–statistical issues in genomewide association studies. N Engl J Med, 357(5), 436-439.
28. Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics, 25(6), 714-721.Barrett, J. C., & Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. Nature genetics, 38(6), 659-662.
29. Hao, K., Chudin, E., McElwee, J., & Schadt, E. E. (2009). Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC genetics, 10(1), 1.
30. Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38.8 (2006): 904-909.
31. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. The American Journal of Human Genetics, 90(1), 7-24.

32. Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. Annual review of genomics and human genetics, 10, 387.
33. Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science, 273(5281), 1516-1517.
34. Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., ... & Todd, J. A. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145), 661-678.
35. Witte, J. S., Gauderman, W. J., & Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. American Journal of Epidemiology, 149(8), 693-705.
36. Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992). Selection of controls in case-control studies: III. Design options. American journal of epidemiology, 135(9), 1042-1050.
37. Ziegler, A., König, I. R., & Thompson, J. R. (2008). Biostatistical aspects of genome-wide association studies. Biometrical Journal, 50(1), 8-28.
38. Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. The American journal of human genetics, 86(1), 6-22.
39. Zuvich, R. L., Armstrong, L. L., Bielinski, S. J., Bradford, Y., Carlson, C. S., Crawford, D. C., ... & Hayes, M. G. (2011). Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. Genetic epidemiology, 35(8), 887-898.
40. Witte, J. S. (2010). Genome-wide association studies and beyond. Annual review of public health, 31, 9.
41. Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic epidemiology, 32(4), 381-385.

42. Rzhetsky, A., Wajngurt, D., Park, N., & Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. Proceedings of the National Academy of Sciences, 104(28), 11694-11699.
43. Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440-9445.
44. Wacholder, S., Chanock, S., Garcia-Closas, M., & Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. Journal of the National Cancer Institute, 96(6), 434-442.
45. Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., ... & Brooks, L. D. (2007). Replicating genotype–phenotype associations. Nature, 447(7145), 655-660.
46. Jannot, A. S., Ehret, G., & Perneger, T. (2015). P< 5× 10− 8 has emerged as a standard of statistical significance for genome-wide association studies. Journal of clinical epidemiology, 68(4), 460-465.
47. Tsepilov, Y. A., Ried, J. S., Strauch, K., Grallert, H., Van Duijn, C. M., Axenovich, T. I., & Aulchenko, Y. S. (2013). Development and application of genomic control methods for genome-wide association studies using non-additive models. PloS one, 8(12), e81431.
48. Devlin, B., & Roeder, K. (1999). Genomic control for association studies. Biometrics, 55(4), 997-1004.
49. International Multiple Sclerosis Genetics Consortium. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. N engl J med, 2007(357), 851-862.
50. Oksenberg, J. R., & Baranzini, S. E. (2010). Multiple sclerosis genetics—is the glass half full, or half empty?. Nature Reviews Neurology, 6(8), 429-437.
51. Fu, J., Festen, E. A., & Wijmenga, C. (2011). Multi-ethnic studies in complex traits. Human molecular genetics, ddr386.
52. Nutting, W. B. (1983). Biology and pathology of hair follicle mites (Demodicidae). Cutaneous Infestations of Man and Animal, ed LC Parsih, WB Nutting, RM Schwartsman, 181-199.

53. Mueller, R. S., & Bettenay, S. V. (1999). An unusual presentation of canine demodicosis caused by a long-bodied Demodex mite in a Lakeland Terrier. Australian Veterinary Practitioner, 29(3), 128-+.

54. Chesney, C. J. (1999). Short form of Demodex species mite in the dog: occurrence and measurements. Journal of small animal practice, 40(2), 58-61.

55. Conroy, J. D., Healey, M. C., & Bane, A. G. (1982). New Demodex sp. infesting a cat: a case report [Demodectic mange]. Journal-American Animal Hospital Association (USA).

56. McDougal, B. J., & Novak, C. P. (1986). Feline demodicosis caused by unnamed Demodex mite. The Compendium on continuing education for the practicing veterinarian (USA).

57. Chesney, C. J. (1987). An unusual species of demodex mite in a cat. The Veterinary record, 123(26-27), 671-673.

58. Chee, J. H., Kwon, J. K., Cho, H. S., Cho, K. O., Lee, Y. J., Abd El-Aty, A. M., & Shin, S. S. (2008). A survey of ectoparasite infestations in stray dogs of Gwang-ju City, Republic of Korea. The Korean journal of parasitology, 46(1), 23-27.

59. Rodriguez-Vivas, R. I., Ortega-Pacheco, A., Rosado-Aguilar, J. A., & Bolio, G. M. E. (2003). Factors affecting the prevalence of mange-mite infestations in stray dogs of Yucatan, Mexico. Veterinary parasitology, 115(1), 61-65.

60. Sischo, W. M., Ihrke, P. J., & Franti, C. E. (1989). Regional distribution of ten common skin diseases in dogs. Journal of the American Veterinary Medical Association, 195(6), 752-756.

61. Scott, D. W., & Paradis, M. (1990). A survey of canine and feline skin disorders seen in a university practice: Small Animal Clinic, University of Montreal, Saint-Hyacinthe, Quebec (1987-1988). The Canadian Veterinary Journal, 31(12), 830.

62. Muller, G. H., Kirk, R. W., & Scott, D. W. (2001). Parasitic skin diseases. Muller & Kirk’s Small Animal Dermatology, 6th edn. Philadelphia, PA: WB Saunders, 457-74.

63. Gortel, K. (2006). Update on canine demodicosis. Veterinary Clinics of North America: Small Animal Practice, 36(1), 229-241.

64. Plant, J. D., Lund, E. M., & Yang, M. (2011). A case–control study of the risk factors for canine juvenile-onset generalized demodicosis in the USA. Veterinary dermatology, 22(1), 95-99.
65. Ravera, I., Altet, L., Francino, O., Bardagí, M., Sánchez, A., & Ferrer, L. (2011). Development of a real-time PCR to detect Demodex canis DNA in different tissue samples. Parasitology research, 108(2), 305-308.
66. Ghubash, R. (2006). Parasitic miticidal therapy. Clinical techniques in small animal practice, 21(3), 135-144.

67. Fourie, L. J., Kok, D. J., Du Plessis, A., & Rugg, D. (2007). Efficacy of a novel formulation of metaflumizone plus amitraz for the treatment of demodectic mange in dogs. Veterinary parasitology, 150(3), 268-274.

68. Muller, G. H., Kirk, R. W., Scott, D. W., Giriffin, C.E. (2001). Canine Demodicosis. Muller & Kirk’s Small Animal Dermatology, 6th edn. Philadelphia, PA: WB Saunders, 457-74.

69. Mederle, N., Darabu, G., Oprescu, I., Morariu, S., Ilie, M., Indre, D., & Mederle, O. (2010). Diagnosis of canine demodicosis. Sci. Parasitol, 11(1), 20-23.
70. Miller, W. H., Griffin, C. E., Campbell, K. L., & Muller, G. H. (2013). Muller and Kirk's Small Animal Dermatology7: Muller and Kirk's Small Animal Dermatology. Elsevier Health Sciences.
71. Caswell, J. L., Yager, J. A., Parker, W. M., & Moore, P. F. (1997). A prospective study of the immunophenotype and temporal changes in the histologic lesions of canine demodicosis. Veterinary Pathology Online, 34(4), 279-287.
72. Krawiec, D. R., & Gaafar, S. M. (1980). Studies on the immunology of canine demodicosis [Demodex canis]. Journal-American Animal Hospital Association (USA).
73. Akilov, O. E., Butov, Y. S., & Mumcuoglu, K. Y. (2005). A clinico-pathological approach to the classification of human demodicosis. JDDG: Journal der Deutschen Dermatologischen Gesellschaft, 3(8), 607-614.

74. Ginel, P. (1996). Demodicose beim Hund.[Demodicosis in dogs]. Waltham Focus, 6, 2-7.
75. Bienvenu, A. L., Gonzalez-Rey, E., & Picot, S. (2010). Apoptosis induced by parasitic diseases. Parasites & vectors, 3(1), 1.
76. Singh, S. K., Dimri, U., Sharma, M. C., Swarup, D., Sharma, B., Pandey, H. O., & Kumari, P. (2011). The role of apoptosis in immunosuppression of dogs with demodicosis. Veterinary immunology and immunopathology, 144(3), 487-492.
77. Singh, S. K., Dimri, U., Sharma, M. C., Sharma, B., & Saxena, M. (2010). Determination of CD4+ and CD8+ T cells in the peripheral blood of dogs with demodicosis. Parasitology, 137(13), 1921-1924.
78. Owen, L. N. (1972). Demodectic mange in dogs immunosuppressed with antilymphocyte serum. Transplantation, 13(6), 616.
79. Healey, M. C., & Gaafar, S. M. (1977). Immunodeficiency in canine demodectic mange. I. Experimental production of lesions using antilymphocyte serum. Veterinary Parasitology, 3(2), 121-131.
80. Grandi, F., Pasternak, A., & Beserra, H. E. O. (2013). Digit loss due to Demodex spp. infestation in a dog: clinical and pathological features. Open veterinary journal, 3(1), 53-55.

81. Ferrer, L., Ravera, I., & Silbermayr, K. (2014). Immunology and pathogenesis of canine demodicosis. Veterinary dermatology, 25(5), 427-e65.
82. It, V., Barrientos, L., Lopez Gappa, J., Posik, D., Diaz, S., Golijow, C., & Giovambattista, G. (2010). Association of canine juvenile generalized demodicosis with the dog leukocyte antigen system. Tissue Antigens, 76(1), 67-70.
83. Ulutas, B., Ural, K., & Ulutas, P. A. (2011). Acute phase response with special reference to C-reactive protein in dogs with generalized demodicosis. Acta Scientiae Veterinariae, 39(3), 980.
84. Felix, A. O. C., Guiot, E. G., Stein, M., Felix, S. R., Silva, E. F., & Nobre, M. O. (2013). Comparison of systemic interleukin 10 concentrations in healthy dogs and

those suffering from recurring and first time Demodex canis infestations. Veterinary parasitology, 193(1), 312-315.

85. Ravera, I., Ferreira, D., Gallego, L. S., Bardagí, M., & Ferrer, L. (2015). Serum detection of IgG antibodies against Demodex canis by western blot in healthy dogs and dogs with juvenile generalized demodicosis. Research in veterinary science, 101, 161-164.

86. Reddy, B. S., & Sivajothi, S. (2016). CD4+ and CD8+ T cells in the peripheral blood of dogs affected with generalised demodicosis. Comparative Clinical Pathology, 25(2), 295-297.

87. Mueller, R. S. (2004). Treatment protocols for demodicosis: an evidence-based review. Veterinary Dermatology, 15(2), 75-89.

88. Živičnjak, T. (2005). A retrospective evaluation of efficiency in therapy for generalized canine demodicosis. Veterinarski arhiv, 75(4), 303-310.

89. Mueller, R. S., Bensignor, E., Ferrer, L., Holm, B., Lemarie, S., Paradis, M., & Shipstone, M. A. (2012). Treatment of demodicosis in dogs: 2011 clinical practice guidelines. Veterinary dermatology, 23(2), 86-e21.

90. Cury, G. M. M., Pereira, S. T., Botoni, L. S., de Oliveira Pereira, R. D., da Costa Telles, T., Ferreira, A. P. L., & Costa-Val, A. P. (2013). Diagnosis of canine demodicosis: comparative study between hair plucking and adhesive tape tests. Revista Brasileira de Ciência Veterinária, 20(3).

91. Pereira, D. T., Castro, L. J. M., Centenaro, V. B., Amaral, A. S., Krause, A., & Schmidt, C. (2015). Skin impression with acetate tape in Demodex canis and Scarcoptes scabiei var. vulpes diagnosis. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, 67(1), 49-54.

92. Sutter, N. B., Eberle, M. A., Parker, H. G., Pullar, B. J., Kirkness, E. F., Kruglyak, L., & Ostrander, E. A. (2004). Extensive and breed-specific linkage disequilibrium in Canis familiaris. Genome research, 14(12), 2388-2396.

93. Meurs, K. M., Mauceli, E., Lahmers, S., Acland, G. M., White, S. N., & Lindblad-Toh, K. (2010). Genome-wide association identifies a deletion in the 3′

untranslated region of striatin in a canine model of arrhythmogenic right ventricular cardiomyopathy. Human genetics, 128(3), 315-324.

94. Awano, T., Johnson, G. S., Wade, C. M., Katz, M. L., Johnson, G. C., Taylor, J. F., ... & March, P. A. (2009). Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. Proceedings of the National Academy of Sciences, 106(8), 2794-2799.

95. Andersson, L. (2009). Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. Genetica, 136(2), 341-349.

96. Tsai, K. L., Noorai, R. E., Starr-Moss, A. N., Quignon, P., Rinz, C. J., Ostrander, E. A., ... & Clark, L. A. (2012). Genome-wide association studies for multiple diseases of the German Shepherd Dog. Mammalian genome, 23(1-2), 203-211.

97. Wolf, Z. T., Brand, H. A., Shaffer, J. R., Leslie, E. J., Arzi, B., Willet, C. E., ... & Wang, X. (2015). Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate. PLoS Genet, 11(3), e1005059

98. Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., ... & Mauceli, E. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature, 438(7069), 803-819.

99. Machiela, M. J., & Chanock, S. J. (2014). GWAS is going to the dogs. Genome biology, 15(3), 1.

100. Ostrander, E. A. (2012). Both ends of the leash—The human links to good dogs with bad genes. New England Journal of Medicine, 367(7), 636-646.

101. Tang, R., Noh, H.J., Wang, D., Sigurdsson, S., Swofford, R., Perloski, M., Duxbury, M., Patterson, E.E., Albright, J., Castelhano, M. and Auton, A., 2014. Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. *Genome biology*, *15*(3), p.1.

102. J Grady, B., & D Ritchie, M. (2011). Statistical optimization of pharmacogenomics association studies: key considerations from study design to analysis. *Current Pharmacogenomics and Personalized Medicine (Formerly Current Pharmacogenomics)*, *9*(1), 41-66.

103. https://genome.ucsc.edu/

104. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559-575.

105. Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics, 23(10), 1294-1296.

106. Agresti, A. (1990) Categorical data analysis. New York: Wiley. Pages 59–66.

107. Agresti, A. (2002) Categorical data analysis. Second edition. New York: Wiley. Pages 91–101.

108. Fisher, R. A. (1935) The logic of inductive inference. Journal of the Royal Statistical Society Series A 98, 39–54.

109. Fisher, R. A. (1962) Confidence limits for a cross-product ratio. Australian Journal of Statistics 4, 41.

110. Fisher, R. A. (1970) Statistical Methods for Research Workers. Oliver & Boyd.

111. Mehta, C. R. and Patel, N. R. (1986) Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered r*c contingency tables. ACM Transactions on Mathematical Software, 12, 154–161.

112. Clarkson, D. B., Fan, Y. and Joe, H. (1993) A Remark on Algorithm 643: FEXACT: An Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. ACM Transactions on Mathematical Software, 19, 484–488.

113. Patefield, W. M. (1981) Algorithm AS159. An efficient method of generating r x c tables with given row and column totals. Applied Statistics 30, 91–97.

114. D. J. Best & D. E. Roberts (1975), Algorithm AS 89: The Upper Tail Probabilities of Spearman's rho. Applied Statistics, **24**, 377–379.

115. Myles Hollander & Douglas A. Wolfe (1973), Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 185–194 (Kendall and Spearman tests).

116. Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

117. Cailliez, F. (1983) The analytical solution of the additive constant problem. Psychometrika **48**, 343–349.

118. Cox, T. F. and Cox, M. A. A. (2001) Multidimensional Scaling. Second edition. Chapman and Hall.

119. Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**, 325–328.

120. Krzanowski, W. J. and Marriott, F. H. C. (1994) Multivariate Analysis. Part I. Distributions, Ordination and Inference. London: Edward Arnold. (Especially pp. 108–111.)

121. Mardia, K.V. (1978) Some properties of classical multidimensional scaling. Communications on Statistics – Theory and Methods, **A7**, 1233–41.

122. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). Chapter 14 of Multivariate Analysis, London: Academic Press.

123. Seber, G. A. F. (1984). Multivariate Observations. New York: Wiley.

124. Torgerson, W. S. (1958). Theory and Methods of Scaling. New York: Wiley.

125. Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics **21**, 768–769.

126. Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics **28**, 100–108.

127. Lloyd, S. P. (1957, 1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory **28**, 128–137.

128. MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, **1**, pp. 281–297. Berkeley, CA: University of California Press.

129. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genome-wide

pedigree-based quantitative trait loci association analysis. Genetics. 2007 177(1):577-85.

130. Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. PLoS ONE. 2007 Dec 5;2(12):e1274

131. Thompson EA, Shaw RG (1990) Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics 46, 399-413.

132. Svischeva G, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nature Genetics. 2012 44:1166-1170. doi:10.1038/ng.2410

133. Belonogova NM, Svishcheva GR, van Duijn CM, Aulchenko YS, Axenovich TI (2013) Region-Based Association Analysis of Human Quantitative Traits in Related Individuals. PLoS ONE 8(6): e65395. doi:10.1371/journal.pone.0065395

134. Lars Ronnegard, Xia Shen and Moudud Alam (2010). hglm: A Package for Fitting Hierarchical Generalized Linear Models. The R Journal, **2**(2), 20-28.

135. Youngjo Lee, John A Nelder and Yudi Pawitan (2006) Generalized Linear Models with Random Effect: a unified analysis via h-likelihood. Chapman and Hall/CRC.

136. Xia Shen, Moudud Alam, Freddy Fikse and Lars Ronnegard (2013). A novel generalized ridge regression method for quantitative genetics. Genetics 193(4), ?1255-1268.

137. Moudud Alam, Lars Ronnegard, Xia Shen (2014). Fitting conditional and simultaneous autoregressive spatial models in hglm. Submitted.

138. Woojoo Lee and Youngjo Lee (2012). Modifications of REML algorithm for hglms. Statistics and Computing **22**, 959-966.

139. Chen WM, Abecasis GR. Family-based association tests for genome-wide association scans. Am J Hum Genet. 2007 Nov;81(5):913-26.

140. Jia, P., Wang, L., Meltzer, H. Y., & Zhao, Z. (2010). Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. Schizophrenia research, 122(1), 38-42.

141. Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I., & Xiong, M. (2010). Genome-wide gene and pathway analysis. European Journal of Human Genetics, 18(9), 1045-1053.

142. Giacomelli, L., & Covani, U. (2010). Bioinformatics and data mining studies in oral genomics and proteomics: new trends and challenges. The open dentistry journal, 4(1).

143. Elbers, C. C., van Eijk, K. R., Franke, L., Mulder, F., van der Schouw, Y. T., Wijmenga, C., & Onland-Moret, N. C. (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genetic epidemiology, 33(5), 419-431.

144. Lesnick, T. G., Papapetropoulos, S., Mash, D. C., Ffrench-Mullen, J., Shehadeh, L., De Andrade, M., ... & Maraganore, D. M. (2007). A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet, 3(6), e98.

145. Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., ... & Amos, C. I. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. European Journal of Human Genetics, 18(1), 111-117.

146. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C., Craddock, N. and Wellcome Trust Case-Control Consortium, 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. The American Journal of Human Genetics, 85(1), pp.13-24.

147. Chen, L., Zhang, L., Zhao, Y., Xu, L., Shang, Y., Wang, Q., ... & Li, X. (2009). Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. Bioinformatics, 25(2), 237-242.

148. Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics, 178(3), 1709-1723.

149. Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... & Kresovich, S. (2006). A unified mixed-model method for association

mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208.

150. Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., ... & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, *42*(4), 355-360.

151. http://pantherdb.org/geneListAnalysis.do

152. Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'Er, I., ... & Olincy, A. (2009). Common variants on chromosome 6p22. 1 are associated with schizophrenia. Nature, 460(7256), 753-757.

153. Bertram, L., & Tanzi, R. E. (2009). Genome-wide association studies in Alzheimer's disease. Human molecular genetics, 18(R2), R137-R145.

154. Ding, K., & Kullo, I. J. (2009). Genome-wide association studies for atherosclerotic vascular disease and its risk factors. Circulation: Cardiovascular Genetics, 2(1), 63-72.

155. Jin, L., Zuo, X. Y., Su, W. Y., Zhao, X. L., Yuan, M. Q., Han, L. Z., ... & Rao, S. Q. (2014). Pathway-based analysis tools for complex diseases: a review. Genomics, proteomics & bioinformatics, 12(5), 210-220.

156. Ohi, K., Hashimoto, R., Ikeda, M., Yamamori, H., Yasuda, Y., Fujimoto, M., ... & Iwase, M. (2015). Glutamate networks implicate cognitive impairments in schizophrenia: genome-wide association studies of 52 cognitive phenotypes. Schizophrenia bulletin, 41(4), 909-918.

157. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. Genome Res. 2010; 20(9):1191–1197

158. Glinsky GV. SNP-guided microRNA maps (MirMaps) of 16 common human disorders identify a clinically accessible therapy reversing transcriptional aberrations of nuclear import and inflammasome pathways. Cell Cycle. 2008; 7(22):3564–3576.

159. http://www.genecards.org/cgi-bin/carddisp.pl?gene=CYP2C18

160. http://www.genecards.org/cgi-bin/carddisp.pl?gene=GOT1

161. http://www.genecards.org/cgi-bin/carddisp.pl?gene=ABCC2

162. http://www.genecards.org/cgi-bin/carddisp.pl?gene=KCNIP2

References for figures

Figure 1.2) https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

Figure 1.3) Atanasovska, B., Kumar, V., Fu, J., Wijmenga, C., & Hofker, M. H. (2015). GWAS as a driver of gene discovery in cardiometabolic diseases. *Trends in Endocrinology & Metabolism*, *26*(12), 722-732.

Figure 1.4) Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95-108

Figure 1.5) Hardin, M., & Silverman, E. K. (2014). Chronic obstructive pulmonary disease genetics: a review of the past and a look into the future. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *1*(1), 33-46

Figure 1.6) Östensson, M. (2012). *Statistical Methods for Genome Wide Association Studies*. Chalmers University of Technology

Figure 1.7) https://www.boundless.com/biology/textbooks/boundless-biology

Figure 1.8) http://www.discoveryandinnovation.com/BIOL202/notes/lecture25.html

Figure1.9) http://www.slideshare.net/AustralianBioinformatics/jian-yang-mixed-linear-model-analyses-of-human-complex-traits-using-snp-data-37071391

Figure 1.10) 6Oksenberg, J. R., & Baranzini, S. E. (2010). Multiple sclerosis genetics—is the glass half full, or half empty?. *Nature Reviews Neurology*, *6*(8), 429-437.

Figure 1.11) http://www.theskinvet.net/veterinary-surgeons/canine-demodicosis/

Figure 1.12) Chesney, C. J. (1999). Short form of Demodex species mite in the dog: occurrence and measurements. Journal of small animal practice, 40(2), 58-61

Figure 1.13) 1Rutan, J. Treating canine demodicosis

Figure 1.14) http://www.willows.uk.net/en-GB/specialist-services/pet-health-information/dermatology/canine-demodicosis

Figure 2.1) Ferrer, L., Ravera, I., & Silbermayr, K. (2014). Immunology and pathogenesis of canine demodicosis. Veterinary dermatology, 25(5), 427-e65

Figure 3.1) Hong, Huixiao, et al. "Pitfall of genome-wide association studies: Sources of inconsistency in genotypes and their effects." (2012).