Identifying Envirogenomic Signatures for Predicting the Clinical Outcomes of Crohn's Disease

Author

Nasir, Bushra Farah

Published

2013

Thesis Type

Thesis (PhD Doctorate)

School

School of Medical Sciences

DOI

10.25904/1912/1046

Identifying Envirogenomic Signatures for Predicting the Clinical Outcomes of Crohn's Disease

Bushra Farah Nasir

M. Medical Research (Biomedical Science)

School of Medical Sciences

Griffith Institute of Health and Medical Research

Genomics Research Centre

Griffith University

Submitted in fulfilment for the Degree of Doctor of Philosophy

April 2013

ABSTRACT

A complex interplay between genetic susceptibility, environmental factors and clinical indicators seem to cause the development of Crohn's disease (CD). Few disorders in clinical medicine are associated with as much chronic morbidity as CD. Genetic factors are the predetermined cause, whereas non-genetic factors seem to further trigger the development of CD. From epidemiological data, based on concordance statistics in family studies, via linkage analysis to Genome Wide Association Studies (GWAS) and Whole Genome Analysis (WGA), robust evidence have been gathered, implicating distinct genomic loci involved in genetic susceptibility to CD.

Most recently, a meta-analysis has been able to implicate 71 distinct genomic loci that seem to be associated with CD development. A study published in the American Journal of Human Genetics has also recently identified more than 200 genes associated with CD, which is more than what have been found for any other disease so far. Monozygotic twins show approximately 50-60% disease concordance, with much lower rates in dizygotic twins (~10%), highlighting the role of both environmental and genetic components in the development of CD.

In earlier decades, CD patients suffered a lack of effective treatment preferences, and patients with moderate to severe CD were often consigned to prolonged systemic corticosteroid therapy and surgery as their only options. From the late 1990s onward, there has been a significant improvement in CD treatment therapy, specifically the widespread adoption of maintenance of immunomodulaters and the advent of biologic agents. The understanding of the natural history of those patients at-risk of severe and complicated CD patient subgroups, particularly an early identification of

patients with the potential for severe disease and its associated complications, would ultimately determine an optimal clinical approach that incorporates appropriate risk-benefit assessment for disease modifying therapy.

The advent of chips for high-throughput genotyping of single nucleotide polymorphisms (SNPs) in 2004 provided researchers with a powerful new tool for unlocking the genetic basis of complex human diseases. GWAS using SNP chips have since identified risk loci for many complex disease traits, though the amount of disease heritability explained by any single SNP has been typically very minor (odds ratios < 1.5). Whilst the SNP chip technologies supporting GWAS are now scientifically well established (more than 1 million SNPs can currently be typed rapidly in a single assay), the bioinformatics methods for GWAS are still underdeveloped and require more empirical experimentation with model traits before widespread application is possible. In particular, for accurately identifying disease risk factors, there is an urgent need to move beyond the single SNP analysis, and into the statistical assessment of multiple SNPs acting independently and/or interactively with other environmental and clinical risk factors.

Association studies are a major tool for identifying genes conferring susceptibility to complex disorders. These traits and diseases are being termed as “complex” because both genetic and environmental factors contribute to the susceptibility of risk development. Extensive experiments in genetic studies for many complex disorders have confirmed that many different genetic variants control disease risk, with each variant having only a subtle effect. Therefore genomic data needs to be integrated with patient environmental and clinical information so that it can be interpreted by trained clinicians who can provide personalised predictive profiles for patients, thus allowing for a more specifically tailored treatment regimen. Furthermore, this will

allow clinical researchers to translate novel disease biomarkers and drug responses into improved treatments, which in turn can be tailored to each patient's individual genome. It is anticipated that such developments would be readily applicable to CD along with other multifaceted diseases.

This research project adapts the basic GWAS design towards identifying envirogenomic signatures of disease, with a major focus on CD. By combining clinical, environmental and genetic factors and analysing them collectively, "envirogenomic signatures" can be created. Through the use of a multi-factor analysis procedure, it is expected that a greater proportion of the disease variance would be explained and therefore the coupled factor set would be of greater predictive value. This study would be a step forward in identifying envirogenomic signatures that may predict clinical outcomes in CD patients with an improved accuracy.

CD associated SNPs play an important part in disease progression and severity, and information from clinical outcomes can be used to identify disease risk prospects for patients, when used with other clinical patient history data. A replicated multi-factorial model was developed and analysed in this section. This model was able to illustrate an 82% risk of requiring surgery for patients diagnosed with both the NOD2 gene and smoking variables. When clinical and/or environmental predictors and genetic predictors are combined to create envirogenomic signatures, high association to clinical severity of disease outcomes is demonstrated (Chapter Four) and this can potentially assist in creating individual risk profiles, useful for subsequent accurate treatment and early diagnosis of disease severity.

This project (Chapter Five) also involved the development of a novel Genomic Signature Analysis (GSA) bioinformatics method which is a comprehensive, step-wise approach for analysing genomic data in order to produce a precise genomic signature for predicting the level of disease risk. Originally outlaid in 2009 by Lea *et al*, further analyses has been able to provide progressive developments of the GSA method. With the inclusion of some modifications, the GSA process has been further improved. A 7-SNP genetic signature was identified using the GSA method by using the Wellcome Trust Case Control Consortium (WTCCC) CD dataset. The completed model was able to achieve an 86% risk of CD for those patients with the set of 7 SNPs identified from the GSA. The GSA method outlined in this study demonstrated a significantly important process to help classify high risk patients of CD and empower clinicians with the ability to diagnose such patients accurately and provide personalised treatment therapies accordingly.

The field of medicine today is rapidly accelerating from inefficient and experimental practices to data-driven and practical clinical applications. Very soon, treatments, diagnosis, prognosis and disease prevention would be custom-made to each individual's specific genotypic and phenotypic dataset. In the near future we anticipate that, through the incorporation of genomic research innovations, personalized medicine advances would no doubt transform the practice of medicine, thereby altering the global healthcare industry and with time, would pave the way for a lengthier and healthier quality of life.

ACKNOWLEDGEMENTS

Throughout my academic career I have been blessed with countless opportunities that have come my way, making my career path more achievable. I am extremely grateful to all those who have walked along with me on every step of this journey, those who I am blessed enough to be able to take for granted. The past three years of my Doctoral research work have been extremely fulfilling and full of cooperation from many people in my life, some of whom I would like to appreciate here. It goes without saying, that I am eternally grateful to God for providing me this opportunity in life and bestowing me with the capabilities to accomplish this task.

First and foremost, my supervisors Dr Rodney Lea and Dr Lyn Griffiths cannot be accredited enough, for all their support and guidance that they have provided so unassumingly. Making special considerations for me to attend meetings with my children, at my home and through teleconferences, I greatly appreciate the kindness and respect that they have always endowed upon me. Without these two people, I am genuinely positive that I would not have been able to complete such an immense undertaking as that is required for a Doctoral degree. I am sincerely thankful to both of these mentors for giving me the guidance and space to learn, so that I could really call this work my own.

The past few years of my academic learning would also not have been possible without the abundant love, support and encouragement of my friends and family. I am extremely thankful to have been blessed with parents who have always encouraged me to excel and push my limits and always provided me with immense educational and emotional care. I am grateful to my mother, Tasneem Akhtar Nasir for being with me and my family every step of the way, and looking after us always.

I am thankful to my father, Dr M. Aslam Nasir for providing inspiration and knowledge, without which I could not have fulfilled my desire to excel in knowledge and develop my career.

I am also deeply appreciative that I have been blessed with a life partner who is both understanding and supportive. When others would have not encouraged me to attempt more in my life, my husband's unquestionable love and reassurance have allowed me to achieve what seemed quite impossible to undertake in the beginning.

I would also like to acknowledge the rest of family, my brothers and their families, and my friends who have been with me every step of the way and provided much encouragement and support whenever needed.

A special note of gratitude is also extended to Amanda Miotto from the eLearning Support Specialist team of Griffith University for her constant assistance through all the computational/technical aspects of this project. I would also like to thank Donia Macartney-Coxsen, David Hall and Miles Benton as part of the Genomics Research Centre lab for providing guidance, feedback and support throughout my candidature. I would also like to extend my appreciation to the Griffith University support services for their support throughout this journey, especially to those from the Research Computing Services. I would also like to appreciate the Australian Government for my APA scholarship, whereby making this research financially possible.

Lastly, I would like to dedicate this thesis to my three delightful children Sarmad, Fauzia and Aleeza, who have made this journey all the more fulfilling, and have always been a constant source of unconditional love and happiness throughout life.

STATEMENT OF ORIGINALITY

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

(Signed)_____

Bushra Farah Nasir

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Anti-TNF/TNF-α: Anti–tumour necrosis factor

AR: Attributable risk

AUC: Area under the curve

AZA: (Imuran) azathioprine

CD: Crohn's disease

C.I.: Confidence interval

dbGaP: Database of Genotypes and Phenotypes

Dx: Diagnosis

GEWIS: Genome-environment-wide interaction studies

GI: Gastrointestinal tract

GWAS: Genome wide association studies

IBD: Inflammatory bowel disease

MR: Misclassification rate

NPV: Negative predictive value

NZ: New Zealand

OR: Odds ratio

PD: Perianal disease

PPV: Positive predictive value

RR: Relative risk

ROC: Receiver operator characteristic (curve)

S.E.: Standard error

TPMT: Thiopurine S-methyltransferase

UC: Ulcerative colitis

WGAS: Whole genome association studies

WGS: Whole genome sequencing

WTCCC: Wellcome Trust Case Control Consortium

QIMR: Queensland Institute of Medical Research

6-MP: Purinethol: 6-mercaptopurine

LIST OF PUBLICATIONS & CONFERENCE PRESENTATIONS

1 Papers accepted for publication:

(i) **Nasir B**, Griffiths L, Nasir A, Roberts R, Barclay M, Gearry R and Lea R "Perianal disease combined with NOD2 genotype predicts need for IBD-related surgery in CD patients from a population-based cohort" (2013) J Clin Gastroenterol 47 (3): 242-245 DOI: 10.1097/MCG.0b013e318258314d

(ii) **B. Nasir**, R. Lea, R. Gearry, A. Nasir, D. Macartney-Coxsin, D. Hall, L. Griffiths (2011) Envirogenomic Risk Profiling to predict clinical outcomes in CD patients, Abstracts for the 35th Human Genetics Society of Australasia Annual Scientific Meeting, Gold Coast, Australia July 31-Aug. 3, 2011, Twin Res Hum Gen 14 (4): 347—85, DOI: http://dx.doi.org/10.1375/twin.14.4.347

2 Papers in preparation for submission or submitted awaiting outcomes:

(i) **B Nasir**, L Griffiths, L Simms, GR Smith, R Gearry, P Bampton, JM Andrews, IC Lawrence, A Nasir, R Lea "An analysis of CD risk factors that lead to an increased risk of surgery: a multivariate cohort analysis"

(ii) **B Nasir,** R Lea, L Griffiths, A Miotto, D Hall, D Macartney-Coxsin, A Nasir "Personalised profiling to accurately predict clinical outcomes in CD using the new Genomic Signature Analysis method"

(iii) **Nasir BF,** Griffith LR, Nasir A, Roberts R, Barclay M, Gearry RB, Lea RA. **(**2013). An envirogenomic signature is associated with risk of IBD-related surgery in a population based Crohn's disease cohort. J

Gastrointest Surg 17 (9): 1643-50, DOI: 10.1007/s11605-013-2250-1, Epub 2013 Jul 2

3 Conference Presentations:

(i) Poster presentation at the Australian Society for Medical Research Conference 2010, "Identifying a genomic signature for predicting the risk of CD" **B. Nasir**, Rod Lea, Lyn Griffiths

(ii) Podium presentation at the Gold Coast Health & Medical Research Conference 2010, "Perianal disease combined with NOD2 genotype predicts need for IBD-related surgery in CD patients from a population-based cohort", **BF Nasir**, LR Griffiths, MA Nasir , RL Roberts, ML Barclay , RB Gearry, RA Lea

(iii) Podium presentation at the Biomarker Discovery Conference 2010, "Identifying envirogenomic risk profiles for predicting clinical outcomes in CD", **BF Nasir**, LR Griffiths, RL Roberts, R Gearry, RA Lea, MA Nasir

(iv) Podium presentation at the Human Genetics Society of Australasia Annual Scientific Meeting 2011, "Envirogenomic Risk Profiling to predict clinical outcomes in CD patients", **B. Nasir**, R. Lea, R. Gearry, A. Nasir, D. Macartney-Coxsin, D. Hall, L. Griffiths

(v) Poster presentation at Australian Society for Medical Research Conference 2011, "Identifying envirogenomic risk profiles for predicting clinical outcomes in CD", **BF Nasir**, LR Griffiths, RL Roberts, R Gearry, RA Lea, MA Nasir

(vi) Podium presentation at the Human Genetics Society of Australasia 2011, "Identifying envirogenomic risk profiles for predicting clinical outcomes in CD", **BF Nasir**, LR Griffiths, RL Roberts, R Gearry, RA Lea, MA Nasir

(vii) Podium presentation at the Biomarker Discovery Conference 2012, "Identifying envirogenomic risk profiles for predicting clinical outcomes in CD", **BF Nasir**, LR Griffiths, RL Roberts, R Gearry, RA Lea, MA Nasir

(viii) Poster presentation at the Joint Conference of Human Genome Meeting and the 21st International Congress of Genetics 2013, Singapore, "Identifying envirogenomic risk profiles for predicting clinical outcomes in CD", **BF Nasir**, LR Griffiths, RL Roberts, R Gearry, RA Lea, MA Nasir

4 Awards:

(i) Prize for the Early Career Researcher talks at the Biomarker Discovery Conference 2010

CHAPTER ONE: AIMS AND SIGNIFICANCE

This doctorate research project investigates and identifies novel envirogenomic signatures for the accurate prediction of clinical outcomes in CD patients. The research carried out as part of this project involved the analysis of a number of different datasets, so that an overall perspective of disease risk prediction could be studied.

The first chapter of this thesis focuses on illustrating the relevance and necessity of this study. The second chapter provides a brief background of the disease itself and describes a comprehensive literature review of CD. Because the scope of the study is two-fold i.e. it performs genetic as well as non-genetic data analysis methods, the third chapter briefly explains some key concepts relevant to this study with the hope that readers from various backgrounds would be able to understand the research presented in subsequent chapters. These concepts are just a brief overview, and by no means a thorough description of the concepts presented in this thesis. Subsequent chapters explain the methods applied throughout this research project and the results obtained are then thereby discussed.

1.1 Overview

CD is a chronic relapsing and remitting inflammatory condition of the intestine. CD and ulcerative colitis (UC) are the two main components of Inflammatory Bowel Disease (IBD) and although the cause of the disease is still unknown, there appears to be a deregulated host immune response to intestinal microbiota in genetically susceptible individuals.

Aetiologically, CD results from a complex combination of genetic, environmental and clinical determinants. The model of disease inheritance is complex and IBD most probably represents a group of heterogeneous disorders which share some but not all susceptibility genes (Noomen *et al*, 2009). Substantial progress has been made in unravelling the genetic background of IBDs; how they interact with environmental factors, each other and the immune system is largely still under investigation (Noomen *et al*, 2009). Even though many environmental factors, and the way that these factors effect individuals with CD, have been identified as associated with CD, precise verification that any single environmental factor is the specific cause of developing CD for any one individual is still not comprehendible. Diarrhoea, abdominal pain, fatigue, weight loss and fever are common features of CD (Hart and Ng, 2011). Clinical traits of CD depend on the location and behaviour of the disease in the Gastrointestinal (GI) tract (Hart and Ng, 2011).

The incidence of CD varies between the various regions of the world. The highest annual incidence of CD was 12.7 per 100,000 people in Europe, 5.0 per 100,000 people in Asia and the Middle East, and 20.2 per 100,000 people in North America (Molodecky *et al*, 2012). Rates vary between 0.1 and 16/100,000 inhabitants with the highest incidence recorded in Northern and Western Europe and North America, while lower rates are recorded in Africa, South America and Asia (Hart & Ng, 2011). The incidence of CD is increasing with time and in different regions around the world (Molodecky *et al*, 2012). A growing number of incidences are occurring in developing countries as well, though the disease is still being recorded with higher incidences in developed countries (Hart & Ng, 2011). Recent highest reported prevalence values for IBD were from Europe with 322 per 100,000 persons having

CD, and North America with 319 per 100,000 persons having CD (Molodecky *et al*, 2012).

The prevalence of IBDs in Australia is 360 per 100,000 individuals, the second highest after Canada (Colgan, 2006). The Geelong Incidence Study in 2008 reported 75 new cases of IBD, out of which 46 CD cases were being evaluated. The peak incidence occurred between the ages of 20 and 24 years, with 42 females and 33 males in this study. The IBD incidence rate of 29.3 per 100,000 is the second highest ever reported in the literature, and the CD incidence rate of 17.8 per 100,000 is the highest ever reported in this study (Wilson *et al*, 2010).

Because its incidence and prevalence are rising in all ethnic groups and because of the systemic nature of the illness, CD concerns an increasingly diverse group of clinicians (Baumgart *et al*, 2012). The past few decades have witnessed an alarmingly high increase in the incidence of IBDs, proliferated from a number of influencing factors that influence lifestyle choices made by individuals. These certain lifestyle trends that are found in a majority of Western nations, are spreading throughout the world including those areas where CD was not abundant before, and therefore as a result, are causing the proliferation of this disease in many diverse regions of the world.

In economic terms, CD exacts a substantial toll: the disease has direct and indirect annual costs estimated at $826 million in the United States alone (Williams *et al*, 2008). A more current evaluation would probably cause this estimated figure to be on the rise. The direct costs include inpatient and outpatient care, surgery, diagnostic and follow-up procedures and laboratory tests, and medication. The indirect costs

include lower productivity such as sick leave, unemployment, and disability (Nurmi *et al*, 2013).

In addition to the costs of medical and surgical therapy, the costs of missed work-hours are considerably high, because CD often strikes people during their most productive working years. One analysis estimated that the proportion of patients with CD who are not capable of full-time work is 25% (Colombel, 2007).

Specifically in Australia, the total cost of IBD in 2005 was approximately $267 million, out of which $220 million was related to medical costs, and $40 million to lost wages (Colgan, 2006). CD is related to significant morbidity, work loss, and impaired quality of life but has little effect on mortality, thus causing a high economic burden to the patient and to society (Gibson *et al*, 2008). CD exerts a significant burden on healthcare expenditures in all countries in which it is widespread because diagnosis is often delayed due to its similarities in symptoms with UC and other GI tract diseases. Besides the economic impacts of the disease on society, a patient's way of life is immensely afflicted with many other burdens such as social and emotional hindrances.

Since 1991, the prevalence of CD has increased by approximately 31% (Schreiber *et al*, 2009) and investigative research has established a bimodal distribution of age at diagnosis, with a large peak in incidence between the ages of 20 and 30, and a second, smaller peak that characteristically occurs between the ages of 60 and 70. During the periods of active disease, CD patients live through increased morbidity and a diminished quality of life. Even after diagnosis is established, it is quite complex to predict the course of the disease and high-risk patients are often under-

diagnosed by clinicians who tend to refrain from prescribing vigorous treatments immediately upon diagnosis.

Patients with CD are a heterogeneous group and consequently the individual course of the disease is difficult to predict: some have a rather mild form, but others suffer frequent flare-ups requesting resection, with symptoms recurring soon after surgery (Bernell *et al*, 2000). Although most patients present with uncomplicated disease at the time of diagnosis, a great proportion of them will either present chronic inflammatory activity despite conventional therapy or develop penetrating complications, intestinal stenosis or perianal disease within the first five to ten years after CD diagnosis (Oger *et al* 2001, Blain *et al* 2002, Nion-Larmurier *et al* 2006). Studies have suggested that three out of four patients with CD will undergo an intestinal resection; and half of them will ultimately relapse (Bernell *et al*, 2000). Half of all patients experience intestinal complications within 20 years after diagnosis (Review, 2011). Due to the chronic nature of the disease, only 12% of patients with CD have been reported to have experienced a relapse-free course 10 years after diagnosis (Bernell *et al*, 2000).

Usually, the disease is monitored for its severity and patient treatment regimens are altered according to the development and location of the disease as the disease progresses. Research over the past decade has also verified the evolving temperament of CD over time, which is characterized by tissue remodelling in areas of chronic inflammation. At 20 years post diagnosis of CD, the rates of inflammatory, stricturing, and penetrating disease are 12%, 18%, and 70%, respectively Rutgeerts *et al*, 2009). In addition, the likelihood of needing surgical resection of the colon at 15 years after CD diagnosis is more or less 70% (Rutgeerts *et al*, 2009). The majority of surgical intervention in CD is not curative, as literature

studies have shown the high incidence of post-operative recurrence of CD in many patients. A first resection will need a second resection within 15 years subsequent of diagnosis (Rutgeerts *et al*, 2009). This clinically significant recurrence of CD post-operatively also implies that many patients will require ongoing medical therapy to manage their disease in the post-operative time period. It is also needless to say that surgical interventions lead to many limitations other than those of a physical nature, and tend to impact those immediate people such as family and close friends who look after the patient as well.

As associated CD SNPs and the genes they modify are identified, tasks including evaluating how risk alleles interact with environmental factors, and with each other, will be of key importance. These tasks may identify environmental triggers in those at risk of severe disease and help elucidate disease pathways. Of major importance for genetic translation will be developing materials for evaluating therapies and disease prevention (Brant, 2013).

Acknowledging the complexity of the disease, current treatment methods are based on disease severity which is difficult to determine using current diagnostic methods. As a result, the task of prescribing an accurate and appropriate medication regimen from the vast range of medications available for treatment can be puzzling. The differences in the clinical, environmental and genetic characteristics of patients mean that they tend to respond differently to the currently available treatment options as well. Therefore, deciding on the most effective drug treatment for an individual patient, or in other words personalizing treatment, can be a real challenge for gastroenterologists and clinicians. This highlights the need for better prognostic tests for predicting clinical outcomes for CD patients.

The significant economic, financial, and health burden of CD, all bring to light the necessity for early intervention to prevent the development of complicated CD. Once disease has progressed to a complicated severe state the need for surgery turns out to be highly essential for most cases; however surgical intervention is not a cure for the disease. In the majority of cases relapse of the development of disease is highly probable. It is essential, therefore, to identify those patients who have a higher risk of developing more complicated disease forms as early as possible. If a patient's level of risk can be evaluated at a very early stage during diagnosis, the progression of disease into complicated deteriorating disease forms can be avoided and eventually the need for surgery can also be eliminated from patient treatment regimens.

In less than two decades, CD genetics has revolutionized our understanding of CD etiopathogenesis. For those still suffering from CD and those with a possible future risk, the outlook is steadily improving. As genetic investigators complete the latter phases of disease variant discovery and increasingly focus research efforts toward the determination of associated variants, the functional alterations that actually result in CD risk are being discovered (Brant, 2013). This research study undertakes the task of identifying CD patients that are of high risk to surgery using envirogenomic risk profiling. It also aimed to identify a multi-factor genomic signature for CD.

1.2 Project Motivation

The current era of research mainly focuses on identifying and discovering novel mechanisms and/or entities in the human body that cause disease states. This field of research is a time-consuming, repetitive and meticulous method of understanding disease pathology. This research purely aims to create methods that will be able to stratify high risk patients quickly and conveniently so that accurate treatment can be provided at the time of diagnosis.

By taking into consideration what is already discovered and known about the pathology of this disease, it is hoped that by analysing this known information, it would become beneficial for predicting those patients at risk of complicated disease, at the time of diagnosis. Grasping this concept and keeping in mind that disease prevention and patient treatment is the ultimate goal of all scientific medical research, the utilisation of factors that influence common complex traits of disease into a clinical tool for the early prediction of complicated disease risk, is the central purpose and challenge for this research project.

While therapeutic treatments that change the natural progression of disease history are being developed, it is more beneficial for the patient if prospective disease risk factors can be identified in an individual prior to their manifestation so that early aggressive treatment approaches can minimise patient suffering and increase patient quality of life. Previous studies have identified risk factors that involve clinical, environmental or genetic factors independently. To have the ability to predict the course and complications of CD based on these risk factors has become significantly important since the development of drugs that can alter the natural progression of CD. Therefore, it makes sense to identify an overall disease risk prediction so that patients can benefit from specific, accurate and most appropriate treatment medications.

In this study, a novel approach is undertaken by taking into consideration a collective model that will include clinical, environmental and genetic risk factors in *combination*. An improved predictive possibility for the risk of complicated disease behaviour that leads to severe clinical outcomes would be identified. It is also hoped that a genomic signature would also be developed based on methodologies adapted from a GWAS approach.

1.3 Research Aims

This research aims to identify envirogenomic signatures for predicting the risk of CD using new bioinformatics methods. It is hypothesized that genomic and environmental factors acting in combination would form personalized profiles for the accurate prediction of clinical outcomes in CD patients. This research also aims to identify a multi-gene genomic signature for CD using genomic profiling methodologies.

Another additional goal of this research study is to create a clinically useful computational tool that could aid gastroenterologists and/or clinicians in diagnosing the onset of CD and predicting the risk of potential clinical outcomes for their patients. As a result of early intervention, accurate diagnosis, and optimal treatment, this tool will eventually reduce the risk of patients developing progressive, complicated, and severe disease and subsequently the need for surgery. This project is divided into two main sections, where each section tackles different methods of predicting risk profiles in complicated and severe CD patients. The first section focuses on analysing genetic, environmental and clinical factors in combination to predict the risk of surgical interventions for CD patients using a step-wise multi-factorial data reduction method. The second section focuses on a comprehensive GWAS-based data reduction method to identify a genetic signature for CD patients.

1.4 Research Hypothesis

This investigative research project will hope to determine if genomic, environmental and clinical factors are able to form personalized profiles that can be used for the accurate prediction of clinical outcomes in CD patients. It is hypothesized that when genomic, environmental, and clinical factors are systematically analysed in

combination, the subsequently developed risk profiles will be of significantly higher predictive power and diagnostic value then when any of the predicting factors are taken into consideration independently.

It is also hoped that by validating the Genomic Signature Analysis (GSA) bioinformatics method, genomic signature profiles can be developed using a GWAS analysis approach. Eventually, the research outcome would lead to the development of a comprehensive computational module for diagnosis in clinical practices. As part of this project a prototype computational tool has been also set up to verify its efficiency in providing accurate diagnostic results. Further evaluations of this tool to evaluate its efficiency in clinical trials are expected in the near future. An overview of the steps undertaken for this research project and its future scope in terms of possible next steps, are outlined in Figure 1.1.

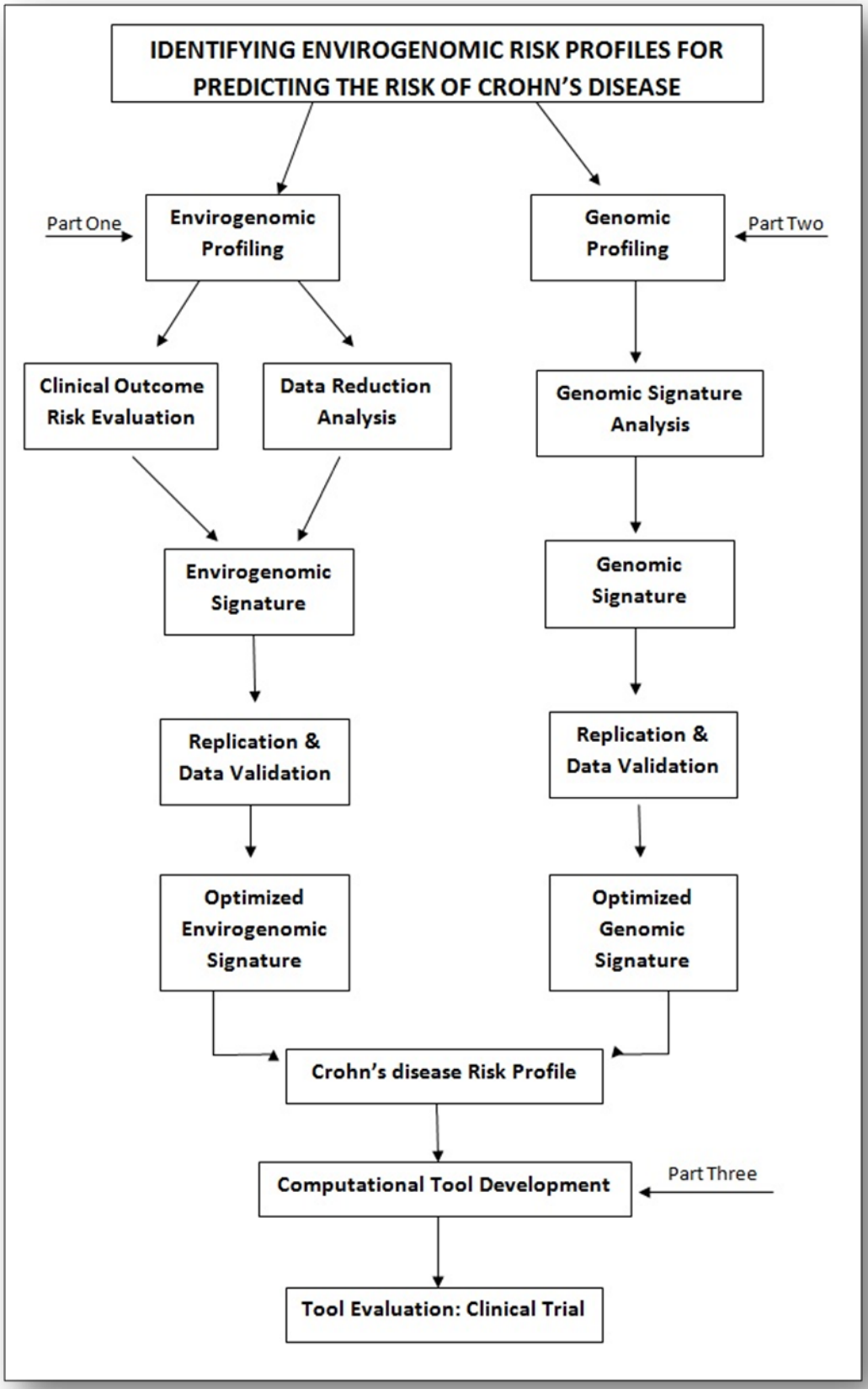

Figure 1.1 The Research Project Flow Chart

1.5 Study Significance

Current approaches for classifying the genes or DNA variants associated with common, genetically influenced human diseases comprise of whole genome comparison of case and control (or test) groups using DNA arrays. By screening DNA markers spanning the entire genome, researchers do not have any need of prior knowledge of the underlying biological foundation of the disease, but rather focus on markers exhibiting allele frequency differences between test groups to distinguish points of interest.

Because of the so many differences in the genetic, clinical, and environmental risk factors associated with the development of complicated and severe CD symptoms, each patient tends to respond differently to the types of treatments that are currently available. This is due to the fact that medications are prescribed based on disease complexity/severity which is difficult to determine based on current diagnostic methodologies. Therefore, gastroenterologists or clinicians face a difficult predicament on deciding the most effective method of treatment for their patients. This emphasizes the need for personalized treatment methods and better prognostic tests for predicting the clinical outcomes of CD and hence limiting disease progression. The ability to identify patients with CD who are at a higher risk for the development and rapid progression of this disease which could lead to surgery, would be invaluable in guiding initial therapeutic choices made by clinicians and/or gastroenterologists at the time of diagnosis.

CHAPTER TWO: BACKGROUND AND LITERATURE REVIEW

2.1 Introduction

CD and UC are the two most common types of IBD. CD is a chronic type of IBD that is characterised by increased transmural inflammation in any part of the GI tract. It is a systematic relapsing inflammatory disease that affects the GI tract with extra intestinal manifestations and associated immune disorders (Baumgart, 2012). The cause of CD remains unidentified. Common symptoms include intense abdominal pain, diarrhoea and weight loss.

The peak age of incidence is usually between the ages of fifteen and thirty. Treatment is usually based on controlling inflammation with drugs (steroids and/or immunomodulaters) or surgery. Due to the sharp rise in incidence during the last third of the twentieth century, IBD now affects up to 1 in 250 of the adult population in the Western world (Stone *et al.*, 2003). Even though the exact aetiology of CD remains unidentifiable, epidemiological data overwhelmingly points to a deregulation of the immune response against the luminal flora in a susceptible host (Limbergen *et al.*, 2009).

CD is very widespread in the Western world, and is increasing in tendency in developing countries as well (Anagnostides *et al.*, 1991). As a result of its rising incidence and prevalence rising in all ethnic groups and because of the systemic nature of the illness, CD is of major concern for an increasingly diverse group of clinicians (Baumgart, 2012).

After the initial characterization of CD in 1932, much still remains to be discovered regarding its aetiology (Economou, 2008). It has a bimodal age of inception, with the

larger peak occurring in the second and third decades and a smaller peak of onset in the elderly; and males and females are equally affected. It is distinguished by frequent attacks of diarrhoea, severe abdominal pain, nausea, fever, chills, weakness, anorexia and weight loss; children with this disease often endure retarded physical development as well (Mosby, 1994). Additionally, there are extra-intestinal manifestations affecting joints, skin, eyes and the liver (Hart & Ng, 2011).

A complex interplay between genetic susceptibility, environmental factors and clinical indicators seem to cause the development of CD. From epidemiological studies, based on concordance data in family studies, via linkage analysis to GWAS and WGAS, robust evidence has been gathered implicating more than 30 distinct genomic loci involved in genetic susceptibility to CD (Limbergen *et al.*, 2009). Since then, a meta-analysis has been able to double this amount, and has potentially implicated 71 distinct genomic loci that seem to be associated with CD development (Noomen *et al.*, 2010). A study published in the American Journal of Human Genetics recently has also identified a novel method of identifying and mapping gene locations for complex inherited diseases like CD. This study has potentially identified more than 200 genes associated with CD, which has been more than that have been found for any other disease so far (Science Daily, 2012).

Monozygotic twins show approximately 50-60% disease concordance, with much lower rates in dizygotic twins (~10%), highlighting the role of both environmental as well as genetic components in the development of CD (Lewis *et al.*, 2007). A positive family history is the most important risk factor for CD; compared with the population prevalence, the relative risk of sibling for a CD patient is 13-36% (Török *et al.*, 2006). The mode of inheritance however remains under debate, with early

segregation studies suggesting a simple Mendelian model with a major recessive gene for at least a proportion of patients with CD (Török *et al.*, 2006).

Over time, the genetic root of this disease is being slowly clarified and it appears to be multigenic (Stenson, 2003). Even though the cause of CD is unknown, an infectious, or environmental, or at times drug related cause in a patient that has a distorted immune inflammatory reaction is generally witnessed first before genetic associations can be deciphered.

The diagnosis of CD is based on clinical signs, x-ray studies using a contrast medium, and endoscopy (Mosby, 1994). Patients with poor prognostic features appear to benefit from early treatment with immunomodulaters drugs and/or anti-tumour necrosis factor therapy (Hart & Ng, 2011). Laproscopic surgery is an occurrence in a majority of patients with complicated disease, although often with disease relapse afterwards. To improve a patient's quality of life prevention strategies should be developed with an increased focus on treating disease symptoms before the disease progresses into complicated and more severe levels. To achieve this, healing patients should not focus on treating symptoms after they appear, and gradually monitoring the disease as it progresses. Instead, risk factors for each diagnosed individual should be identified prior to their manifestation, preferable at the first indication of CD development.

Despite significant research which has unravelled the genetic background of CD, it is also important to identify the clinical and environmental risk factors that lead to a progressive disease course in order to create a thorough personalized patient profile. This would allow for advancement in personalized medicine and effective treatment therapies could be generated. By taking into consideration a patient's individual

personalized profile, it can be possible for clinicians to prescribe a more specific disease management approach so that appropriate treatment can be administered before severe disease progression and complicated disease outcomes are developed.

Those patients at risk of developing complicated disease forms should be treated with aggressive practices of therapy at the time of diagnosis to prevent the development of disease into complex acute stages. In order to identify causative risk factors for CD, genetic factors should not be studied alone, but environmental and clinical factors should also be taken into consideration when analysing individual patient disease characteristics. This would allow the development of a complete, thorough and precise personalized predictive profile for each individual patient, which would ultimately be of immense benefit for diagnostic purposes.

2.2 An Historical Account of CD

In 1932, Crohn, Ginzburg and Oppenheimer described thirteen patients with 'regional ileitis'; before their publications, numerous others had reported various cases of what were in retrospect possibly CD as well (Koltun, 2007). Medical historians advocate that CD may have been first described as early as 1682 to 1771; reports of diseases indicative of CD have occurred in 1806, 1813, 1828, 1875, 1907, 1908, 1909, and 1913 (Chiodini, 1989). The difficulty in distinguishing the colonic form of CD and UC confused the diagnosis and treatment of these illnesses until their differences were clarified by classic publications by Brooke in 1959, and Lockhart-Mummery and Morson in 1960 (Koltun, 2007).

2.3 CD Subtypes

To distinguish the various subtypes of CD from each other, the disease characteristics that have more predictable courses than the unpredictable natural history of CD taken as a whole, are examined. They are classified according to their area of symptom presentation, and can be commonly classified into five subtypes as described below by Health Information Publications (2013).

***Gastroduodenal CD*:** Gastroduodenal CD, which involves the stomach and the duodenum, is often misdiagnosed as ulcer disease. Symptoms of gastro duodenal CD includes loss of appetite, weight loss, nausea, pain in the upper middle of the abdomen, and vomiting.

***Jejunoileitis*:** Jejunoileitis is CD of the jejunum. Symptoms include mild to intense abdominal pain and cramps after meals, diarrhoea, and malnutrition caused by malabsorption of nutrients. Fistulas may form and may increase the risk of developing infections outside of the GI tract.

***Ileitis*:** Ileitis affects the ileum. Symptoms include diarrhoea and cramping or pain in the right lower quadrant and peri umbilical area, especially after meals. Malabsorption of vitamin B12 can lead to tingling in the fingers or toes (peripheral neuropathy). Folate deficiency can hinder the development of red blood cells, putting the patient at higher risk of developing anaemia. Fistulas can develop, as can inflammatory masses.

***Ileocolitis*:** Ileocolitis is the most common type of CD. It affects the ileum and the colon. Often, the diseased area of the colon is continuous with the diseased ileum, and therefore involves the ileocecal valve between the ileum and the colon. In some cases, however, areas of the colon not contiguous with the ileum are involved.

Symptoms of ileocolitis are essentially the same as those present in ileitis. Additionally, weight loss is also common.

***Crohn's Colitis (Granulomatous Colitis)*:** Crohn's colitis affects the colon. It is distinguished from UC in two ways. First, there are often areas of healthy tissue between areas of diseased tissue; UC is always continuous. Second, while UC always affects the rectum and areas of the colon beyond the rectum, Crohn's colitis can spare the rectum, appearing only in the colon.

2.4 Pathophysiology

CD is a chronic transmural inflammation of the GI tract that may involve any part from the mouth to the anus but typically affects the ileum, colon or perianal region. Ileocolitis occurs in about 45% of patients, ileal pattern of involvement is in about 30% and colon pattern of involvement occurs in about 25% of patients (Grover, 2007). Recent studies have also altered this breakdown of disease pathology to patients with ileo-caecal disease at 40%, 30% having exclusive ileal disease and 25% exclusively colonic forms of disease. Perianal involvement occurs in about one-third of patients (Hart & Ng, 2011). The pathogenetic mechanisms in CD are multiple including epithelial damage or abnormal epithelial cells, defective mucus, and acute and chronic inflammation (Anagnostides *et al.*, 1991). Clinical characteristics of CD are probably the most significant predictors of the natural progression of complicated CD outcomes.

2.4.1 Intra-Intestinal Physiology

Typical histological characteristics include patchy transmural inflammation, fissuring ulceration, granulomas, and submucosal lymphoid aggregates. Because the disease

involves transmural inflammation, it appears to skip areas in the GI tract; e.g. a healthy area of small or large bowel can be adjacent to a diseased area. As the disease tends to be discontinuous, giving rise to "skip" lesions, the affected bowel is oedematous and associated with fat wrapping on the serosal surface (Hart & Ng, 2011). Histologically there is predominantly transmural inflammation although this is usually submucosal. Focal patchy chronic inflammation (lymphocytes and plasma cells), focal crypt irregularity (discontinuous crypt distortion) and non-caseating granulomata (not related to crypt injury) are the generally accepted microscopic features that allow a diagnosis of CD (Hart & Ng, 2011).

The initial presentation of CD can usually be characterized as either obstructing or fistulising. Obstructive disease is the result of inflammation narrowing the intestinal lumen and blocking the flow of intestinal contents (Stenson, 2003). Fistulising disease occurs when the inflammatory process stretches completely through the intestinal wall. The escape of bacteria through these defects in the wall can cause abscesses. Extension of the inflammatory progression into neighbouring organs produces fistulae (Stenson, 2003). Obstruction and fistulisation are not mutually exclusive, in fact the enteric end of a fistula is often found immediately upstream from a stricture (Stenson, 2003). Metagenomic research suggests that up to four major bacteria phyla (Bacteoidetes, Firmicutes, Actinobacteria, and Proteobacteria) consisting of thousands of mostly anaerobic species colonise the human gut with steep, stomach-acid driven, proximal-distal gradient (Baumgart, 2012). Species diversity in the gut normally also varies according to temporal, individual, dietary and drug induced factors. However, healthy intestinal microbiotica variation is overall stratified and not continuous (Baumgart, 2012). CD is not caused by diminished commensal diversity alone, but requires a susceptible genotype – as

confirmed by research in mice with human-relevant susceptibility mutations (Baumgart, 2012).

2.4.2 Extra-Intestinal Physiology

CD may present insidiously or acutely and symptoms can vary from vague gastrointestinal upset to severe systemic features of fever, malaise and tachycardia (Hart & Ng, 2011). Previously, it was known that extra intestinal manifestations occur in about 15% and up to 30% in colonic disease and depend on the area of involvement of the disease. However, more recent studies have indicated that extra-intestinal manifestations can be seen in about 25 to 40% of IBD patients (Levine *et al.*, 2011), a figure that has increased over time. Twenty-five percent of IBD patients have more than 1 extra-intestinal manifestation during disease course, and the development of one extra-intestinal manifestation coincides with an increased risk of developing a second extra-intestinal manifestation (Levine *et al.*, 2011). Abdominal pain secondary to partial obstruction is the most frequent presentation (Harrison, 2008). Acute abdominal pain may be confused with appendicitis or *Yersinia ileitis*, but a careful history usually detects previous episodes that have not been acknowledged. Fever, malaise, anorexia, and lassitude usually occur in the active form of the disease. Weight loss alone, without diarrhoea or pain, may be the presenting feature in extensive inflammatory small bowel disease and may be confused with anorexia nervosa in adolescents. Approximately 70-90% of patients would have diarrhoea, 45-66% would have abdominal pain and/or 65-70% would have weight loss. Rectal bleeding is more common in patients with rectal involvement. Obstructive symptoms of nausea/vomiting and abdominal pain/fullness are common in patients with ileal disease, when stricturing occurs. Perianal fistulas are a common complication of CD, occurring in around a third of patients. Some

patients present with perianal fistulas before or at the time of diagnosis of CD. Perianal disease (including skin tags, fissures, anal ulcers, fistulas, abscesses, anorectal strictures) generally denotes a more aggressive CD phenotype (Hart & Ng, 2011).

Other extra-intestinal manifestations affect organs other than the gut, such as the joints, skin, eyes and hepatobiliary system which primary scelorosing cholangitis. The overall incidence of such manifestations is about 30%, with joints being most commonly involved, followed by skin, eye and the hepatobiliary system (Hart & Ng, 2011). These extra intestinal features are described below by Crawford (2005):

***Skin*:** Erythema nodosum is more commonly seen in children with CD 'Metastatic' CD – a rare, nodular, necrotic, granulomatous skin lesions due to a vasculitis – is seen on the limbs and in women under the breasts or on the vulva.

Buccal Mucosa: Aphthous ulceration is more commonly seen in CD.

Eye: Anterior uveitis is seen in 5% of patients with active Crohn's colitis.

Joints: Seronegative polyarthritis usually involves larger joints asymmetrically; it responds when activity of disease diminishes or after colectomy.

Muscle: This is a polymyositis-like syndrome where diseased connective tissue triggers inflammation and muscular weakness.

Secondary Amyloidosis: Nephritic syndrome is the most common presentation, but cardiac, GI, liver and splenic involvement may also occur.

Hypercoagulable State: Elevations levels of factors V and VIII as well as fibrinogen, and lower levels of antithrombin III, have been described; venous thromboses are more common than arterial.

Nutrition: Nutritional deficiencies may account for obscure symptoms, including weakness (Vitamin C, potassium, magnesium), lassitude (iron, B12, folate),

rashes (Niacin, zinc) or altered taste (zinc), but only occur in very extensive disease or after major resection (Travis *et al.*, 2006).

2.5 Diagnostic Techniques

The diagnosis of CD is established on clinical, endoscopic, radiological and histological criteria. This standard remains unchanged in spite of the introduction of new molecular technologies for the analysis of serum proteins and genetic sequences, respectively (Nikolaus and Schreiber, 2007). There is no single definitive diagnostic examination test, so making the correct diagnosis of CD depends on the skills of a clinician to incorporate evidence from patient history, physical discoveries and diagnostic tests. IBD serologic markers (antibodies) may help to differentiate CD from UC and offer prognostic information on disease severity in CD but these markers should not be used to make an initial diagnosis of IBD (Walsh *et al.*, 2011).

The only genetic test approved for use in clinical care of IBD patients is a baseline thiopurine S-methyltransferase (TPMT) enzyme level drawn prior to prescribing the thiopurine analogues, 6-mercaptopurine (6-MP; Purinethol) and azathioprine (AZA; Imuran). Some patients do not have enzyme levels sufficient to metabolize these drugs and can develop hepatotoxicity; it is considered standard of care to confirm sufficient levels before starting these medications (Walsh *et al.*, 2011).

Diagnosis is confirmed by clinical history and examination combined with biochemical, endoscopic, histological and radiological tests. A full medical history should include a history of onset of symptoms, recent travel, family history of IBD, drug history (antibiotics and non-steroidal anti-inflammatory drugs), prior appendectomy and smoking status (Hart and Ng, 2011).A family history of IBD makes the diagnosis more likely (Grover, 2007).

Physical examinations include the individuals' general wellbeing, temperature, blood pressure, heart rate, weight and height, abdominal tenderness or masses, inspection of the oral mucosa and perineum, and a digital rectal examination. The presence of perianal fissures or fistulas and anal induration are suggestive of CD. In patients with mild or moderate CD, physical examination may be normal, whereas those with severe disease will present with fever, tachycardia and abdominal tenderness with palpable inflammatory mass (Hart & Ng, 2011). More specifically, establishing a diagnosis depends on a number of examinations as described below by Travis *et al*. (2006) in detail.

***Sigmoidoscopy and rectal biopsy*:**

- Examines the lining of the lower third of the large intestine (the sigmoid colon).
- Necessary even when the mucosa is macroscopically normal (up to 20% have microscopic granulomas).

Small bowel radiology:

- Radiological tests can include plain x-rays; contrast x-rays, CT scans and multiphase CT enterography.
- Performed first if diarrhoea, pain and weight loss are the presenting features.
- Colonoscopy should subsequently be arranged to exclude Crohn's colitis.

Colonoscopy:

- Examines the lining of the entire large intestine (colon), and sometimes can peek into the very end of the small intestine (or ileum).

- Colonoscopy is preferable to a barium enema if there is diarrhoea or visible rectal bleeding, since aphthoid ulcers are more readily detected and multiple biopsies can be taken.

- Complete small bowel radiology is then advisable, even if the terminal ileum has been demonstrated by reflux or barium, to exclude more proximal disease.

Blood tests:

- Anaemia is common, usually due to iron deficiency rather than B12, or folate deficiency. C-reactive protein (CRP) is the most sensitive inflammatory marker, but an elevated erythocyte sedimentation rate (ESR) or platelet count and a low albumin in a patient with recurrent abdominal pain and weight loss is usually due to CD.

Stool:

- Examination of pathogens and Clostridium difficile toxin assay if diarrhoea is severe.

Once the diagnosis is made it is necessary to establish whether the symptoms are due to active disease or complications. A combination of clinical, blood and imaging tests is needed as no one test is sufficient. Determining the type of CD presenting also depends on a number of characteristics as described below.

Proctitis:

- This presents with bleeding alone, mild increase in stool frequency or constipation, haemorrhoids and solitary rectal ulcer syndrome associated with rectal prolapse can be excluded by proctosigmoidoscopy.

- In older patients (< 40 years), endoscopic examination for colorectal neoplasia is mandatory.

***Irritable bowel syndrome-like*:**

- While altered bowel habit with abdominal pain or discomfort is most commonly caused by IBS, this does not cause fever, weight loss or nocturnal symptoms.

- If these symptoms are present, a colonoscopy and other clinical investigations are recommended.

- The major differential diagnosis of ileocolonic CD is ileal tuberculosis, which is rare in Australia.

Appendicitis-like:

- Most patients who present in this way and have acute ileitis seen during a laparotomy usually do not have CD.

- The cause is usually not known but a minority are infected with Yersinia spp., which can be diagnosed by serological tests.

- For those with CD, presentation usually follows a period of intermittent pain and bowel disturbance.

Chronic diarrhoea without bleeding:

- Among the many causes, of particular note is microscopic colitis, which is characterized by waxing and waning diarrhoea.

- The colon appears normal at colonoscopy, but histological examination shows mucosal inflammation with characteristic features

Extra-intestinal manifestations are highly common in CD patients. An early diagnosis and treatment from medical practitioners is usually recommended to manage the organ systems involved, and subsequently reduce disease progression and the appearance of complications (Levine *et al.*, 2011)

2.5.1 Systems of Classification for IBDs

Over the years, a number of incentives have caused investigators to readdress the complex issues involved in the classification of IBDs. Accurate classification of these diseases would have potential benefits with respect to patient counselling, assessing disease prognosis, and particularly with choosing the most appropriate treatment for each disease subtype (Satsangi *et al.*, 2006).

From the original anatomic classification of CD in 1975, there have been three subsequent published classification systems. One of these, the Vienna Classification, arose from a 1998 World Congress of Gastroenterology Working Party that attempted to prospectively design a simple and objective CD phenotypic classification that encompassed components of age at onset, anatomic location and disease behaviour (Fedorak *et al.*, 2004).

Subsequent application of the Vienna Classification to clinical practice has demonstrated that the CD phenotype changes markedly over time, with nearly 80% of inflammatory disease ultimately evolving into a stricturing or penetrating pattern of behaviour, and 15% undergoing a change in anatomic location. Furthermore, in controlled studies, the ability of IBD experts to similarly identify the disease phenotypes, using the Vienna classification, ranges from poor to fair. Taken together, these failings markedly limit the utilization of the Vienna classification of CD in clinical trials or disease management (Fedorak *et al.*, 2004).

In 2003, a Working Party of investigators with an interest in the issues involved in disease sub-classification was formed with the aim of summarizing recent developments in disease classification and establishing an integrated clinical, molecular, and serological classification of IBD (Satsangi *et al.*, 2006). Although the Vienna classification is still not widely used in clinical practice, researchers have increasingly returned to it and have assessed its applicability and utility. The Montreal revision of the Vienna classification has not changed the three predominant parameters of age at diagnosis, location, and behaviour, but modifications within each of these categories have been made. With respect to age of onset, the Montreal classification allows for early onset of disease to be categorized separately as a new A1 category for those with age of diagnosis at 16 years or younger, whereas A2 and A3 account for ages of diagnosis at 17–40 years and over 40 years, respectively.

Table 2.1 The Montreal and Vienna Classification Systems in Comparison.

	Vienna	**Montreal**
Age of Diagnosis	A1 < 40 yrs. A2 > 40 yrs.	A1 < 16 yrs. A2 17-40 yrs. A3 > 40yrs.
Location	L1 ileal L2 colonic L3 ileo-colonic L4 upper	L1 ileal L2 colonic L3 ileo-colonic L4 isolated upper
Behaviour	B1 non-stricturing non-penetrating B2 stricturing B3 penetrating	B1 non-stricturing non-penetrating B2 stricturing B3 penetrating P peri-anal disease

The differences between the Vienna and the Montreal classifications (Table 2.1) are briefly pointed out below as explained by Satsangi *et al.* in 2006.

> *"With respect to disease location, the major limitation of the Vienna classification was felt to be that each of the four locations described were mutually exclusive. The major difficulty had arisen with the inability of the Vienna classification to allow upper gastrointestinal disease to coexist with more distal disease. As investigations for upper gastrointestinal involvement become more accessible and feasible with the introduction of wireless capsule endoscopy, it is apparent that upper gastrointestinal disease is relatively common, and may coexist with ileal and with colonic disease. Therefore, in the revised Montreal classification these parameters are no longer mutually exclusive. Issues were also identified for integration into the Montreal classification regarding disease behaviour. There are now substantial data that perianal fistulising disease is not necessarily associated with intestinal fistulising disease, and it was felt that perianal disease alone required separate sub-classification. A further issue with regard to classification of disease behaviour is the observation that disease behaviour is dynamic over time. Recent studies have reinforced this, demonstrating that patients with predominantly inflammatory disease at diagnosis are very likely to develop fistulising or stricturing complications within 5, 10, and 20 years".*

2.6 CD Treatment Options

In earlier decades, CD patients suffered a lack of effective treatment preferences, and patients with moderate-to-severe CD were often consigned to prolonged systemic corticosteroid therapy and surgery as their only options. From the late 1990s onward, there has been a significant change in CD treatment therapy, specifically the widespread adoption of maintenance immunomodulaters and the advent of biologic agents (Binion, 2010). Understanding the natural history of these at-risk CD patient subgroups, particularly early identification of patients with the potential for severe disease and its associated complications, will ultimately determine an optimal clinical approach that incorporates appropriate risk-benefit assessment for disease modifying therapy (Binion, 2010).

Over the past two decades, treatment for CD has been based on clinical parameters despite the observations that (1) laboratory parameters can predict the risk of relapse for patients in a clinical remission; and (2) even in the setting of a complete clinical remission, i.e. after an ileo-cecal resection, without maintenance therapy there is a nearly inevitable progression from histologic to endoscopic to clinical recurrence. However, it was not until recent trials with anti–TNF biologic agents that, despite discrepancies between clinical and biologic (endoscopic and C-reactive protein normalization) remissions, the achievement of endoscopic remission had greater impact on long term outcomes, such as the need for hospitalizations and surgeries (Hanauer, 2012).

The introduction, in the last two years, of the innovative M2A (mouth-to-anus) capsule endoscope has made the entire small bowel straightforwardly reachable. Preliminary studies have established that the M2A capsule endoscope can display

lesions of CD when the identification is alleged but other modalities have not been able to validate it (Selby, 2003).

Mucosal healing throughout the progression of CD should not rely on antibiotics or corticosteroid therapy treatments. However, it can be attained with immunomodulaters and biotherapies as treatment forms. But, the therapeutic trepidations then advance from the achievement of clinical remission to the prevention of complications such as fistulas, abscesses and/or stenosis. Many studies have shown that treatment that is more aggressive from the beginning (top-down versus step-up therapy) seems able to achieve mucosal healing more often than the customary treatment technique.

Therapy for IBD has come a long way and is no longer simply the treatment of symptoms. Evolutions in end points for clinical trials and clinical practice goals, along with improved disease foretelling and optimization of medical therapies, provide examples of how this field has taken a lead in the personalization of medicine that will alter the course of these chronic and immune-inflammatory disorders (Hanauer, 2012).

2.6.1 Medical Management of CD

The major goals of therapy are the control of symptoms, induction of remission, healing of endoscopic lesions, and prevention of complications. Physicians have become increasingly eager to utilize traditional immunosuppressive agents as well as innovative biologic therapies (Mahadevan *et al*, 2001). Conventional medications for all types of IBDs are small molecule drugs, most of which were developed for use in other diseases before being found to be effectual for the management of CD. The drugs of foremost significance for inducing remission are steroids (systemic and

controlled release), aminosalicylates and novel biological agents. Medical treatment for CD has transformed extensively in the past years with the surfacing of biologic therapies such as TNF-α that target assorted factors in the inflammatory cascade (Mahadevan *et al*, 2001). The accomplishment of biological therapy is greatly supported by infliximab, which has radically enhanced medical therapy in CD (Stokkers, 2007). Infliximab remains the only confirmed successful TNF-α therapy (Legnani *et al*, 2007).

Over the past decade, several exciting alternative approaches to the medical treatment of this disease have been developed including biologic, probiotic and aphaeresis therapies which have certain advantages over traditional drug therapies. Newer biologic (natalizumab) or cytokine-based therapies (monoclonal antibody to interleukin-6) have shown preliminary evidence of effectiveness in controlled trials, but neither have yet been permitted by the US Food and Drug Administration (or any other Food Authority) and consequently have not been commercialized to date (Legnani *et al*, 2007). However, tacrolimus, a potent calcineurin inhibitor and inhibitor of interleukin-2 expression, has shown efficacy in CD, notwithstanding at the price of considerable potential toxicity (Legnani *et al*, 2007).

Immunomodulaters are principally used for altering the blueprint of the disease. Medical administration principles entail confirming disease action, measuring its rigorousness, considering its locality, and discussing treatment choices. The pharmacologic treatment for CD incorporates a host of immunosuppressive agents:

***Therapy of active disease*:**

- This is usually of short duration (only months) and, as patients are symptomatic, side effects may be more acceptable.

- The choice of drugs depends on severity and site of lesions.

***Corticosteroids*:**

- These are the principle agents for inducing remission of active CD and have rapid therapeutic effects and high efficacy.

- Severe disease requires aggressive intravenous therapy, while for less severe disease the oral route is generally favoured.

- Side effects can be reduced without compromising efficacy by use of corticosteroid enemas, foams or suppositories in colitis effecting the rectum and/or sigmoid and descending colon and by use of oral budesonide coated with pH controlled resin (reduces systemic availability) in ileocaecal CD.

Mesalazine-delivering drugs:

- Oral preparations have efficacy in mild colitis, but their efficacy in ileitis has not been established.

- They have no proven additional benefit if corticosteroids are being used to induce remission.

- The choice of mesalazine-delivering drugs is based largely on cost.

Immunosuppressive agents:

- About 80% of patients tolerate these drugs without side effects, and concern about their safety in long term use is diminishing with wider use and close monitoring.

- Methotrexate has efficacy in CD and can be given orally or intramuscularly, but takes weeks to exhibit maximal effects.

Antibiotics:

- Perineal CD responds poorly to corticosteroids, but imidazoles and ciprofloxacin (alone or in combination) are of benefit.
- They need to be used over months in moderately high doses.
- Early relapse is common when therapy is withdrawn.

Maintenance therapy:

- After control of active disease, maintenance therapy is important to prevent relapse.
- As this therapy is long term and symptoms are minimal, side effects are less acceptable.

Four maintenance strategies are generally applied:

1. Sulfasalazine appears to be of benefit in CD
2. Coated mesalazine reduces the chances of relapse of ileocaecal disease after 'curative' surgical resection by up to 30-40%, and is therefore recommended in this situation
3. Azathioprine also has efficacy in maintaining remission and should be considered in patients with frequent or severe relapse
4. Cessation of smoking, often overlooked, markedly reduces chance and severity of relapse.

From time to time, new drug therapies are developed for treating CD and studies are performed frequently evaluating the efficiency of these forms of treatment. Recently, Ustekinumab, an antibody proven to treat the skin condition psoriasis, has shown positive results in decreasing the debilitating effects of CD (Sandborn *et al.*, 2012).

One third of patients with moderate-to-severe CD do not respond to current treatments with TNF inhibitors, which regulate the body's immune system and inflammation. Another one third of patients only have a temporary response. Therefore, the biggest challenge in treating patients with CD is managing patients whose bodies are resistant to TNF inhibitors. Ustekinumab blocks two proteins that cause inflammation, IL12 and Il12. This finding is a significant first step towards a new treatment option for CD patients (Sandborn *et al.*, 2012).

2.6.2 Surgical Treatment of CD

Many CD patients persist through the life-span of their disease alternating between various forms of medical therapies, hoping to determine a specific treatment regimen that would be the most effective in accurately treating the symptoms of CD that they develop. A large number of patients however, despite long periods of medical treatments, eventually require the need for surgical interventions during the duration of their disease.

Although medical management may offer symptomatic relief primarily, its long term results are inadequate (Hultén, 1988). Surgery is reserved for complications of disease, or for severe limited disease unresponsive to medical therapy. Alternating research suggests that surgery in CD has clear indications and anywhere between 50% of patients require surgery in their lifetime to approximately 80% of CD

patients necessitating surgery at some point during their disease course (Walsh *et al.*, 2011).

Surgery becomes mandatory for the majority of patients, often due to the advancement of intestinal obstruction, intra-abdominal abscesses, or internal or external fistulas (Hultén, 1988). Intestinal resection is more likely to be required for ileal CD than for colonic disease. Approximately 50% of all patients with CD will require at least one intestinal resection within 10 years of diagnosis and 40% will require a further operation within 10 years of the first. Multiple extensive small bowel operations can result in short bowel syndrome and intestinal failure, and limited resections or strictureplasty should be performed where possible. A laproscopic approach is preferred for ileo-colonic resections in CD where appropriate expertise is available (Hart & Ng, 2011).

Delaying surgery in preference of continued aggressive medical therapy despite a poor reaction may lead to avoidable suffering, an increased chance of drug side effects and extra perilous or complicated surgery in the future. However surgery is not curative, postoperative relapse rates are incredibly high and failure of the small intestine may compromise nutrition. Incidence-based studies have shown that three out of four patients with CD will ultimately require intestinal resection, and half of these patients will relapse clinically (Bernell *et al.*, 2000).

Prevention of postoperative CD recurrence is an important goal in management. Evidence suggests that disease severity prior to surgery is a predictor of recurrence. High-risk postsurgical patients may need to start or continue anti-TNF or immunomodulater therapy. In others, endoscopic or radiographic visualization of the

surgical anastomosis 3 to 6 months after surgery may help determine whether to recommence medical treatment (Walsh *et al.*, 2011).

Surgery for small bowel disease should be conservative or 'minimal', generally involving limited resection of extensively involved intestine and, less commonly, strictureplasty of short, well defined strictures. Minimal surgery for the colon is still notorious as an alternative for surgery; total colectomy with ileostomy or ileorectal anastomosis, segmental resection, and non-functioning ileostomy are all existing proposed options. Ileoanal pouch surgery has mediocre results in CD and should be avoided.

The second European evidence-based consensus on the treatment and management of CD has also suggested that the current general trend is to only perform resection on the part of the intestine that is causing symptoms. They determined that surgery is indicated in the event of localised ileocaecal disease with obstruction and without inflammatory signs (C-reactive protein). In case of active CD with abscess, the initial use of antibiotics and percutaneous or surgical drainage is recommended, followed by secondary resection if necessary. In case of small bowel disease, stricturoplasty is an alternative to resection. It is traditionally used for strictures less than 10 cm. Good results have, however, been reported with nonconventional stricturoplasties, and they are worth trying in order to avoid short bowel syndrome. Although feasible, colonic stricturoplasties are not recommended due to the increased risk of cancerization. In cases of localised colon disease (less than one-third of the organ), only the diseased part should be surgically removed. If there is involvement of both extremities of the colon in CD, two segmental resections can be considered. If short stenoses are accessible, endoscopic dilation is preferable. The safety and efficacy of this method

was established mainly for anastomotic stenosis after ileocolic resection (Eugene, 2011).

The benefits of surgery as a result of CD on the quality of life are depicted within many studies including those by Thirlby (2001). Prevention of postoperative CD recurrence is an important goal in its management. Evidence suggests that disease severity prior to surgery is a predictor of recurrence. High risk post-surgical patients may need to start or continue anti-TNF or immune-modulator therapy. In others endoscopic or radiographic visualization of the surgical anastomosis three to six months after surgery may help determine whether to recommence medical treatment (Walsh *et al.*, 2011). During the past three decades the evidence has been accumulating in favour of a minimally invasive approach to IBD.

CD is probably one of the most challenging diseases to treat laproscopically for colorectal surgeons, especially when the disease is located in the colon and involves multiple segments, thus explaining the fact that in the United States the majority of CD patients are still approached with open surgery. Current data suggest a shorter length of stay, shorter ileus, faster recovery and less postoperative pain, along with minimally invasive surgery. On the other hand, significantly longer operative times with laparoscopy are universally reported. Overall, the goal and responsibility is to explore new avenues for a true minimally invasive approach to treating CD (Zoccali and Fichera, 2012).

2.6.3 Health Maintenance for CD

CD patients are at risk of osteopenia and osteoporosis, with some research estimating prevalence as high as 70%. This risk is attributable to several reasons, including antecedent corticosteroid use, vitamin D and calcium malabsorption, and the

osteoporotic consequences of chronic inflammation. Standard guidelines from the American Gastroenterological Association (AGA) recommend that the following high-risk CD patients should be screened for osteoporosis: those with a history of vertebral fractures, postmenopausal females, males older than 50 years, those on chronic corticosteroid therapy, or those with hypogonadism. Patients with osteoporosis should begin treatment with a bisphosphonate and calcium supplementation (Walsh *et al.*, 2011).

The nutrition and diet of a patient with CD plays a crucial role and are thoroughly related (Kelly, 2008). Nutritional therapy also produces no drug-induced side effects and is successful in inducing and sustaining remission in CD (Takahashi, 2007). This can occur as a result of reduced food intake, digestion and absorption, enhanced requirements, distorted metabolism of nutrients, amplified losses and drug-nutrient interactions. Patients with IBDs often have iron deficiency due to blood loss and chronic inflammation. Folate and vitamin B12 levels are also important to assess, especially in patients with Anaemia (Walsh *et al.*, 2011).

Malnutrition is frequent in patients with CD; therefore diet has a significant role in re-establishing and sustaining nutritional status. Therapeutic Guidelines Limited (2007) has provided the following recommendations for diet therapy:

- Total parenteral nutrition is indicated if the GI tract is not functional, or if oral or enteral nutrition cannot be tolerated.
- In the majority of patients a normal diet, with or without polymeric oral nutritional supplements, is usually recommended.

- A low-residue diet may assist in controlling diarrhoea and pain related to food intake during an exacerbation.

- It is also useful as longer-term measure in patients with strictures, stenosis or low grade intestinal obstructive symptoms.

- Lactose-intolerance may occur in patients with diffuse disease in the small intestine, but lactase activity should return as remission is achieved.

- Steatorrhoea resulting from extensive ileal disease or resection can be managed by a low-fat diet. If energy requirements cannot be met, supplementation with medium-chain triglyceride oil is recommended.

- Cholestyramine may be helpful in patients with bile salt diarrhoea; however it may worsen steatorrhoea due to bile salt depletion.

- Micronutrient deficiencies should be corrected using appropriate supplementation. Micronutrients particularly at risk include iron, zinc, vitamin B12, calcium, magnesium, folic acid, and vitamin D.

- Any foods that repeatedly exacerbate symptoms should be avoided. However at present there is no evidence for use of exclusion diets in preventing a relapse of CD.

The utilization of total parenteral nutrition or the enteral administration of an elemental diet as a particular therapy is notorious (Stenson, 2003). Elemental diets consist of amino acids, monosaccharides, vitamins, minerals, and essential fatty acids (Stenson, 2003), and supply nutrients in their simplest forms – protein as free amino acids, carbohydrate as glucose or short chain maltodextrins and fat as short chain triglycerides (Takahashi, 2007).

In CD, elemental diets have been used more comprehensively in children than in adults (Stenson, 2003). The mechanism of action of elemental diets is explained by Takahashi (2007):

> *"The therapeutic effect of elemental diets is reported to result initially from a reduction in immune stimuli in the gut as a result of the removal of dietary whole protein. More recently, however, whole protein enteral diets have been shown to be as effective as elemental diets in active CD suggesting that the therapeutic effect of an elemental diet cannot be entirely attributed to the exclusion of whole protein. Given that elemental diet therapy has demonstrated to significantly reduce intestinal permeability continuous intermittent administration may allow maintenance of normal intestinal permeability and thereby protect against relapse."*

Dietary supplementation with fish oil may have anti-inflammatory action as the n-3 fatty acids in fish oil fight in the substrate pool for the enzymes that produce prostaglandins and leukotrienes (Stenson, 2003). Problems with nutritional therapy incorporate a decline in compliance owing to unpalatability, moderation of appetite and social inconvenience (Takahashi, 2007). Further, unfavourable side effects of this treatment include liver damage due to fatty infiltration and acute glucose load, demonstrating that care is required not to overload or raise the amount of an elemental diet too rapidly over time (Takahashi, 2007).

The progression of disease to lead to cancerous indications is also significant for patients with IBDs such as CD. The combined risk of developing colorectal cancer in CD is as high as fourfold compared to the general population. Although there have been no large trials that determine a survival benefit from increased surveillance in

CD, indirect evidence of benefit has led the major gastroenterological societies to recommend regular annual surveillance with colonoscopy in those who have had Crohn's colitis for eight or more years (Walsh *et al.*, 2011).

CD patients are often on immunosuppressive medications and thus are at higher risk for infection as well. CD patients should routinely be vaccinated for the following if they are not yet immune: hepatitis A, hepatitis B, influenza, tetanus, streptococcal pneumonia (*Pneumococcus*), diphtheria, pertussis, and varicella. Meningococcus and human papilloma virus vaccines should be administered to target populations (adolescents and young women, respectively). It is important to administer vaccinations prior to the administration of immunomodulaters, anti-TNF therapy, or steroids, because of the decreased immune response while on these agents (Walsh *et al.*, 2011).

Fertility is also decreased for both men and women with CD. In men, constant therapy with immunomodulaters (MTX, 6-MP) reduces fertility rates. This condition is revocable with termination of therapy. For women, fertility is diminished with ongoing MTX therapy and in those with a history of pelvic surgery. Women making an effort to get pregnant should not be on MTX, as it is an abortifacient and teratogen. Annual gynaecologic check-ups are particularly vital in women with CD, as they are at increased risk for cervical cancer also.

Women with CD incline to improve clinically throughout pregnancy; though, CD increases the risk for preterm birth. Sustained use of other agents during pregnancy, principally the immunomodulaters, is debatable; though they are classified as category C risk medications, it is commonly agreed that the advantage of sustained use in patients who are well controlled on their current regimen overshadows the risk

of disease flare on cessation. Individualized planning and close management of care between the patient's primary provider, a high-risk obstetrician, and an expert gastroenterologist is vital for a fruitful pregnancy and healthy delivery (Walsh *et al.*, 2011).

2.6.4 The Top Down vs. the Step Up Treatment Approach in CD

Determining the best method of treatment approach for any individual with CD is quite complex and varies between each individual. CD is a heterogeneous disease with approximately 80% of patients having a chronic progressive disease leading to complications, surgeries and potentially socio-professional marginalization and the rest having a simple benign history. Recent studies have shown that early management with immunosuppressive treatments and/or TNF-α agent could change the natural history of the disease and avoid the development of severe complicated disease for some individuals. The treatment therapy regime should thus be personalized according to the risk of developing such disabling disease (Louis *et al.*, 2009). As a result, a fundamental question in the treatment of CD is whether early treatment with biologics is a more effective strategy than the use of traditional steroid therapy followed by biologic therapy. Recent clinical data suggest that some patients may benefit from a "top-down" approach or early biologic treatment, whereas others are more suited to a "step-up" approach (Hanauer, 2003).

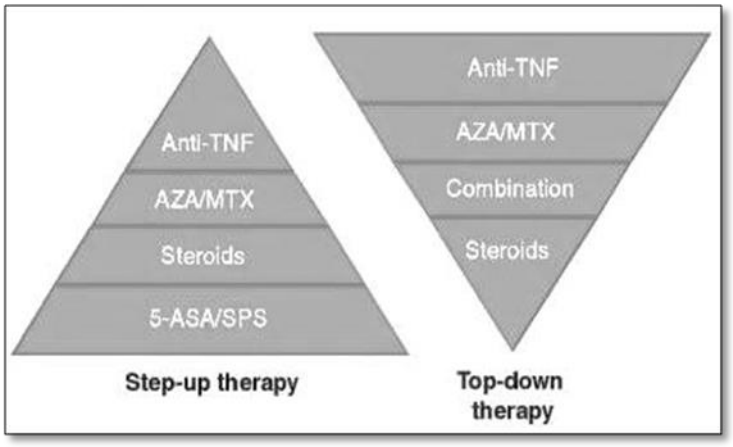

Figure 2.1 The Step Up and Top Down Therapy Approaches

Over the past decades clinical trials have been done to maintain remission based on the categorization of CD severity. The conventional treatment for CD, the step-up approach involves first administering steroids in order to control the patient's symptoms (abdominal pain and bloody diarrhoea); the next step involves administering immune-suppressing drugs, which prepare the body to receive the third medication: an antibody that curbs the inflammatory response at the root of the disease.

The alternative strategy, called top-down therapy, employs early use of immune-suppressing drugs combined with an antibody in order to address the disease from the start. Symptom-treating steroids may never even be needed. The introduction of biologic therapy, and particularly the use of anti-TNF therapy, has provided a powerful tool in the treatment and management of CD (Feagan, 2008). The early introduction of intensive therapies has the aim of avoiding complications and improving quality of life, and is supported by the assumption that these drugs interfere with the natural history of the disease. Hodges (2008) also verifies the benefits of the top down approach by stating that the top down treatment strategy for

newly diagnosed CD appears to be safer and more effective than standard therapy. The top-down approach is also considered to be safer because it spares patients exposure to steroids (Feagan, 2008).

Corticosteroids are very efficient drugs to suppress acute inflammation. The short term outcome (30 days) of a first course of steroids shows that approximately 60% of patients would have a complete response, 30% a partial response and 10-15% would be non-responders. However, after 1 year, only one-third of patients would have a prolonged response. It is clear that corticosteroids are of no benefit in maintaining remission or preventing new flares, and that they do not heal the mucosa (Vermiere *et al.*, 2006). On the other hand, the superiority of immunomodulaters in inducing, but particularly in maintaining remission of CD has been well documented (Vermiere *et al.*, 2006).

The achievement of early and sustained healing due to biotherapies should be able to modify the natural history of the disease, limit the number of hospitalisations and surgical intervention and reduce the incidence of sequelae, as suggested by the second European evidence-based consensus on the diagnosis and management of CD (2011). In view of the side effects of immunomodulaters, biotherapies and their combination, it is necessary to better discern the patients that are prone to benefit from this approach and for what length of time. Indeed, even though these treatments and their combinations in particular are more effective than 5-ASA or corticosteroid therapy, they do still carry potential risks. For the thiopurine drugs, these include an increased risk of lymphoma; for the TNF antagonists, the risk of stimulating latent tuberculosis or opportunistic infections; and for the combination of both drugs, the risk of hepatosplenic T-cell lymphoma, which though exceptional, is particularly serious (Eugene, 2011).

The nineties (1990s) have been characterized by the introduction of biological therapies, designed to block or neutralize pro-inflammatory cytokines, which play a role in the pathogenesis of CD. TNF- α is a pivotal cytokine in CD and increased TNFα levels are detected in serum, tissue and stools of patients with CD. Infliximab, a mouse-human chimeric antibody to TNFα is a very efficacious therapy for the treatment of refractory luminal and fistulising CD. Almost 80% of patients experience a rapid improvement of their symptoms and almost 50% of patients have a complete remission. Infliximab also has a steroid sparing effect. Besides neutralization of TNFα, the long-lasting effects of infliximab have been attributed to complement activation, antibody-dependent cellular cytotoxicity and induction of apoptosis (Vermiere *et al.*, 2006).

Several observations still limit the use of infliximab as first-line treatment in adult CD patients. In particular, the epidemiological observation that over 50% of CD patients have a mild disease over time and will never require aggressive therapies is against the indiscriminate use of top-down strategy. Lack of markers that enable identification of high-risk patients, discussions about long-term safety and the high costs of infliximab are further factors supporting a more careful approach to the management of CD. Hence, the best method of treatment approach still remains unanswered, and clinicians tend to adopt a less-aggressive and economical method of treatment for their patients in order to avoid unnecessary drug and surgical interventions.

2.7 CD Epidemiology

At present the cause of CD is remains unidentified. Consequently epidemiologic data has been gathered over many years in the anticipation of providing some indication

to the aetiology of the disease (Koltun, 2007). The frequency of the disease has been rising worldwide recently, but its spread has been decelerating in vastly affected countries (Karlinger *et al.*, 2000).The countries of greatest incidence and prevalence cover the industrialized world (Stenson, 2003). In geographic areas where the incidence of these types of diseases has been minor it is now escalating (Stenson, 2003). CD is definitely emerging worldwide as a major public health threat, with increasing reports of paediatric disease further underlying this threat. Changes in lifestyle, at the regional, national or international level, seem to play an etiological role in the increasing incidence of this disease (Economou, 2008). CD can be diagnosed at any age but most commonly presented between 10 and 40 years of age, with a smaller peak in the seventh decade. CD is marginally more common in women than in men, and there is a notably high incidence in Ashkenazi Jews (Hart & Ng, 2011).

2.7.1 CD Prevalence

The incidence of CD varies worldwide and is increasing with time and spreading into new geographical regions. The highest annual incidence of CD was 12.7 per 100,000 person-years in Europe, 5.0 person-years in Asia and the Middle East, and 20.2 per 100,000 person-years in North America (Molodecky *et al.*, 2012). Rates vary between 0.1 and 16/100,000 inhabitants, with highest incidence recorded in Northern and Western Europe and North America, with lower rates recorded in Africa, South America and Asia (Hart & Ng, 2011). The number of people with CD has been steadily increasing, particularly among young people and in non-Western societies. Estimates of prevalence vary depending on whether figures are derived from primary, with 200/100,000 persons or secondary/tertiary institutions with 70-100/100,000 persons (Hart &Ng, 2011). Estimates position the frequency of CD at

174–201 cases per 100,000 persons in the United States alone (Lichtenstein *et al.*, 2009). In 2012, the highest reported prevalence values for IBD were from Europe with 322 per 100,000 persons having CD, and North America with 319 per 100,000 persons having CD (Molodecky *et al.*, 2012).

The prevalence of IBD in Australia as recorded in 2010 was 360 per 100,000, which was second highest only to Canada (Wilson *et al*, 2010). The Geelong Incidence Study in 2008 reported that 75 new cases of IBD out of which 46 CD cases were evaluated. The peak incidence occurred between the ages of 20 and 24 years, with 42 females and 33 males in this study. The IBD incidence rate of 29.3 per 100,000 is the second highest ever reported for Australia in the literature, and the CD incidence rate of 17.8 per 100,000 is the highest ever reported from Australia (Wilson *et al.*, 2010).

Geographically, the prevalence of the disease has a gradient from North to South and, to a slighter degree from West to East; the Western-Eastern inconsistency can be accredited to diversity in Western life styles (Karlinger *et al.*, 2000). Rates in central and southern Europe are fairly minor, in South America, Asia and Africa it remains uncommon but appears to be increasing (Stenson, 2003).

CD has been escalating in incidence to a remarkable degree over the past 20-30 years with a two-to-tenfold increase depending on population and region studied. These dramatic increases suggest an environmental effect, because a genetic factor would probably not alter disease rates so rapidly (Koltun, 2007).

2.7.2 Ethnic Occurrences

CD is three to eight times more frequent in Jewish than in non-Jewish persons, however, its occurrence among Israeli Jews is much lower than among American and

European Jews (Stenson, 2003). In the United States, the incidence of CD in the African-American population has been one-fifth to one-half that in the Caucasian population, but in recent years the gap appears to be lessening (Stenson, 2003).

2.7.3 CD Incidence with respect to Gender and Age

According to age, the onset of the disease occurs more frequently in the second or the third decade of life (Karlinger *et al.*, 2000); the peak age of onset for CD is between 15 and 25 years (Stenson, 2003). In several but not all series, a second, lesser peak of incidence occurs between 55 and 65 years (Stenson, 2003). This disease also occurs in childhood, although the incidence is much lower before the age of 15 years than after the age of 15 (Stenson, 2003).

In children CD incidence is higher in males with a switch of the gender ratio incidence at 15 years old. Conversely CD phenotype does not appear different according to gender. These results suggest that hormonal factors could be involved in the susceptibility to CD but not in its phenotypic manifestations (Massouille, 2008). These gender differences of disease distribution may symbolize perplexing environmental aspects such as the use of tobacco as well (Stenson, 2003). Another study has also shown that the diagnosis of CD in older patients is uncommon, and patients diagnosed at age 60 years and older are less likely to develop complicated CD than patients diagnosed at younger ages. The discovery of less complicated diseases in elderly patients may be as a result of differences in disease duration and location compared with younger patients (Quezada *et al.*, 2012).

2.8 CD Risk Factors

A complex interplay between genetic susceptibility, environmental factors and clinical indicators seem to cause the development of CD. Few disorders in clinical medicine are associated with as much chronic morbidity as CD (Shanahan, 2002), however still no definite aetiology has been defined, and the complex nature of CD supports the notion that its origin is likely to be multi-factorial (Mamula *et al.*, 2003). Current theory suggests that in genetically predisposed individuals, environmental factors and maladaptive immune responses to GI flora generate a deregulated inflammatory cascade creating mucosal injury (Mamula *et al.*, 2003). Genetic factors influencing the development of CD is at about 50-60% indicating that the level of hereditary risk for CD is relatively high in comparison to environmental and other clinical factors. A study on the biopsychosocial understanding of IBD was published in 2001 (Ringel *et al.*) and suggested that early in life, genetic and environmental factors determine the susceptibility to IBD. Later, psychosocial factors and modifiers interacting with the physiologic/pathologic states will influence when and how the disease is experienced symptomatically, the individual's illness behaviour, and the clinical outcome.

2.8.1 Clinical Risk Factors

A vast amount of studies have shown that the development of CD is also associated with a number of clinical risk factors that depend on the pathophysiology of symptoms. Patients diagnosed with CD are known to be at an increased risk of bowel cancers and lymphoma. Also, since CD is an autoimmune disease, patients are therefore predisposed to a wider spectrum of cancers (Hemminki *et al.*, 2009). The clinical risk factor of having cancer may also be in line with previous studies

performed on CD patients who have developed GI cancers such as colorectal cancer (Gyde *et al.*, 1980).

Similarly, perianal disease is an indicative symptom of progressive complicated CD (Gearry *et al*, 2004). Established risk factors for perianal disease include colonic disease and young age at disease onset (Ingle *et al*, 2007). The clinical course of CD is more aggressive in patients with perianal involvement (Gearry *et al*, 2004). Although CD location remains relatively stable, behaviour changes over time and perianal disease is a strong predictor of developing more complicated CD according to Gearrys' study in 2004.

Regarding having had an appendectomy and then developing the need for IBD-related surgery, studies have shown that patients who have had previous appendectomies are at a much increased risk of developing CD and those CD patients with a history of perforated appendicitis mostly developed complicated and progressive symptoms of CD (Anderson *et al.*, 2003). However, several studies have shown an inverse relationship between appendectomy and subsequent development of IBD although these findings remain contentious. A study by Singhal *et al.* in 2010 has shown no relationship between appendicitis/appendectomy and the development of IBD to further show the unreliability of this factor as being a consistently significant risk factor for CD.

Earlier studies have correspondingly suggested that appendectomy is associated with a substantially reduced risk of certain types of IBD such as CD, particularly where the underlying diagnosis is acute appendicitis however further analysis of these studies have shown that Coeliac disease and perforated appendicitis are negatively

associated irrespective of the timing of the conditions. Not surprisingly however, CD increases the risk for appendectomy without appendicitis.

It is sufficient to say that the vast array of clinical factors that are associated with CD and have been briefly elucidated upon, all play significant roles in the development of this disease. Along with the other environmental and genetic factors that have also been clarified previously, clinical factors demonstrate a progressive and aggressive impact on the progression of complicated disease.

2.8.2 Environmental Risk Factors

The incomplete concordance rate for CD within monozygotic twins (< 50%) and the variation in risk for some ethnic groups living in diverse geographical locations lend sustenance to the function of environmental factors in the pathogenesis of this disease (Shanahan, 2002). Diet, the role of the early ages, smoking habits and the influence of hormonal status and drugs are viewed as useful contributing factors in the manifestation of this disease (Karlinger *et al.*, 2000).

The prevalence of environmental risk factors for CD are also seen by the striking boost in frequency of CD within the more-developed world over the past 50 years, and the disease's amplified identification with progressive industrialization in less-developed countries. Fundamentals within a shifting environment that might influence growth of the mucosal immune system, the enteric micro-flora, or both, include enhanced sanitation, consumption of sterile and non-fermented foods, vaccination, and age at first exposure to intestinal pathogens (Shanahan, 2002). Other factors such as oral contraceptives, refined sugar, prenatal events, childhood infections, microbial agents, and domestic hygiene have been found to be associated with CD.

Smoking is associated with a three-to-four fold increase in the risk of developing CD (Selby, 2003). It also leads to a more aggressive course with more frequent relapses, more admissions and more time spent in the hospital (Selby, 2003). In another study patients with CD were significantly more likely to be smokers than the controls, and the association was stronger for a smoking habit before the onset of the disease than for a current smoking habit, the relative risks for smokers compared with non-smokers being 4.8 and 3.5 respectively (Somerville *et al*, 1984).

Studies have also shown that smoking is an independent risk factor for clinical, surgical and endoscopic recurrence in CD and also that ex-smokers run a risk of recurrence similar to those of non-smokers, hence indicating that giving up smoking soon after CD-surgery is associated with a lower probability of recurrence (Cottone *et al.*, 1994). On the other hand, another study also indicates the association of ex-smokers at diagnosis as well as post-diagnosis smokers to the risk of IBD-related surgery as clinical outcomes of the disease (Nasir *et al*., 2013). This study by Nasir *et al* (2013) particularly examines a range of detailed smoking factors to come to the conclusion that ex-smokers at diagnosis and post-diagnosis smokers specifically are particularly at a higher risk of surgical interventions. Smokers are more likely to develop CD than non-smokers (as opposed to UC where smoking is protective) and the disease tends to be more difficult to manage in smokers, who appear to need more immunosuppression and surgical intervention (Hart & Ng, 2011).

Baines (2004) described the association of antibiotic consumption to developing CD. This large UK based study found that patients diagnosed with the disease were about a third more likely to have taken antibiotics in the past, compared with controls, indicating that CD patients were 32% more likely to have taken antibiotics than controls. Another prospective study also investigated the relationship between prior

antibiotic use and developing CD and concluded that there was a significant association between these two factors (Card *et al*, 2004).

Another study by Feeney *et al* in 2002 indicated that improved childhood living conditions are associated with increased risk of CD. Overall, these findings strongly supported the assertion that childhood environment is an important determinant of the risk of any type of IBD in later life, with quite distinct risk factors for CD (Feeney *et al*, 2002). Morris *et al* (2001) also implicates in his study findings that there is a significant link between IBD and left handedness which may be genetic and/or environmental in origin.

Although dietary factors are likely to be of key importance in CD, no dietary components have ever been identified that consistently trigger a flare. Excess refined sugar and low intake of fibre have been associated with CD, but dietary manipulation of sugars/fibre has had no demonstrable impact on disease course. Elemental or polymeric diets are beneficial treatments for children and adults with CD and are associated with mucosal healing (Hart & Ng, 2011).

Some data suggest that a diet high in sugars, fat, and meat intake may increase the risk of CD (Walsh *et al.*, 2011). Obesity may also play a role in the pathogenesis of CD and it may be that obesity-related enteropathy is a distinct entity or a sub-type of CD (Mendall *et al.*, 2009). Beliefs of a causal relationship between dietary factors and IBDs persist and the increase in incidence of CD in countries like Japan and South Korea further implicate a Westernized diet as the underlying cause.

Because inflammation occurs in the digestive tract, antigens present in the gut lumen, including components of food and intestinal bacteria, have been implicated in the development of CD. However, no specific dietary component has been identified, suggesting that many foods contain putative contaminants. A Swedish review (2004)

appraised all dietary studies conducted up to the mid-80s and concluded that the methodologies used both in the study design and data analysis made it impossible to infer anything from the results of these earlier studies. Furthermore, studies carried out after the 1980s have not been clear with any improvement in the quality, and at most the improvement has been marginal (Ekbom, 2004).

There may be a role for enteral nutrition in the maintenance of adult CD, but studies suggesting this were small. Omega-3 fatty acids have also been studied in the treatment of CD, due to their anti-inflammatory properties, but a recent Cochrane review on the subject did not reveal any convincing evidence to routinely recommend their use. Lactose intolerance is a very common cause of gastrointestinal symptoms in the general population, and is even more common in patients with IBD. One study reported prevalence in CD of 40%, compared with 29.2% in controls considered "low ethnic risk" for lactose malabsorption. It is reasonable to recommend a lactose-free diet during flares symptomatic of diarrhoea and bloating (Walsh *et al.*, 2011).

The importance of childhood factors in the development of IBDs has also been confirmed in a recent study (Gearry *et al.*, 2009). The duration-response protective association between breast-feeding and subsequent development of IBDs requires further evaluation, as does the protective effect associated with a childhood vegetable garden (Gearry *et al.*, 2009) in this study of CD risk factors.

The roles of these environmental factors that have clearly been established in IBDs promote an understanding of three hypotheses associated with these factors that play a role on the pathophysiology of IBD: the hygiene, infection and cold chain hypotheses (Jantchou *et al*, 2006). Besides the hygiene and infection risk factors that have already been mentioned, the cold chain hypothesis proposes a link between CD

and the ‘cold chain’ which causes chronic infestation of the digestive tract by psychotrophic bacteria. The cold chain hypothesis explains that CD was provoked by infantile exposure to micro-organisms that can survive refrigerator temperature. Findings point to refrigeration as a potential risk factor for CD and furthermore, cold-chain development paralleled the outbreak of CD during the 20th century. The cold chain hypothesis suggests that psychrotrophic bacteria such as *Yersinia spp*. and *Listeria spp*. contribute to the disease (Hugot *et al*, 2003).The cold chain has produced many benefits for Western societies, including the prevention of enteric infections, allowing more people access to a well-balanced diet, and the economic development of agriculture and fishing. These advantages clearly outweigh the putative risks discussed here and, in the absence of experimental evidence, practical conclusions should not be drawn.

Evidence implicating microbiota in the aetiology of CD comes from numerous animal models of IBD which remain healthy when kept in “germ-free” conditions but develop colitis when colonized by commensal microbiota. In the closest analogous situation in humans, when the faecal stream of a patient with CD is diverted, the downstream inflammation resolves but reappears when continuity is restored.

Recent studies in animal models have led to the widespread belief that microbial constituents provide the antigenic stimulus in CD, and the most common hypothesis is that chronic inflammation results from an abnormal host immune response to normal luminal flora (Bulois *et al*, 2008). Many organisms that cause these abnormalities have been suggested, including adherent invasive *Escherichia coli*; *mycobacteria paratuberculosis*; *Listeria; Pseudomonas fluorescens;* and *Bacteroidesvulgatus*. On the other hand, *Faecalibacterium prausnitzii* appears to be

protective (Hart & Ng, 2011). Besides these organisms, early exposure to measles virus, live-attenuated measles vaccination and prenatal exposure to measles were all associated with an increased risk to developing CD (Bulois *et al*, 2008).

There appears to be an imbalance in the microbiota with altered diversity and richness. The impact of the genetic background, smoking and diet on the microbiota is also poorly understood (Hart & Ng, 2011). Despite there being a multitude of studies, microbial agents appear to be intimately involved in the pathogenesis, but no single causative microorganism has been found. Chronic intestinal and systemic inflammation requires the interaction of both genetic and bacterial factors, but neither on their own is sufficient to induce chronic colitis or enteritis. Finally, resolution of the enigma may be hampered owing to disease heterogeneity, and the possibility exists that CD might represent a syndrome with multiple aetiologies (Bulois *et al*, 2008).

A recent study in 2011 (Pugazhendhi *et al*.) also confirmed several known environmental risk factors for CD. They published that CD has been shown to be associated with markers of better childhood hygiene, such as birth order, urban residence, separate bedroom as a child, and the availability of hot water taps in the house. Smoking also influences disease course and severity in CD. In this study other exposures postulated to be associated with CD include the use of fast foods, cola drinks, toothpaste, antibiotics and oral contraceptives.

Throughout literature, some of which has been described here, all studies indicate that CD has a strong component whereby it can be associated with environmental factors that cause disease risk. This understanding makes it logical to analyse and include CD environmental risk factors, along with the other risk factors associated

with this disease, when investigating individual risk towards complicated disease forms.

2.8.3 Hereditary Risk Factors

A positive family history is considered to be the most important risk factor for any type of IBD. Compared with the population prevalence, the relative risk of sibling for a CD patient is 13-36 % (Török *et al.*, 2006). The mode of inheritance however remains under debate, with early segregation studies suggesting a simple Mendelian model with a major recessive gene for at least a proportion of patients with CD (Török *et al.*, 2006). Monozygotic twins show approximately 50-60% disease concordance, with much lower rates in dizygotic twins (~10%), highlighting the role of both environmental and genetic components in the development of CD (Lewis *et al.*, 2007). It is becoming increasingly clear over the last few years that CD is clinically heterogeneous and that there is a genetic basis for the clinical heterogeneity (Stenson, 2003). Not only are patients with CD more likely to have relatives with CD but those relatives are likely to have CD that is similar to that of the proband in terms of anatomic location, age of onset, and disease behaviour (inflammatory vs. fistulising vs. stenotic) (Stenson, 2003).

Genetic studies show that one-fourth of IBD patients have an affected family member but heredity as an etiological factor is stronger in CD than in UC (Karlinger, 2000). The lifetime risk of developing CD among first-degree relatives of affected individuals is 8.9% for offspring, 8.8% for siblings, and 3.5% for parents (Stenson, 2003). Although the relatives of CD patients are more likely to have CD, at an incidence of 30 to 100 times that of the general population, they are also at a greater risk for incidence of UC than the general population as well (Stenson, 2003).

There is a thirteen times greater prevalence of CD in first-degree relatives in comparison with non-relatives (Tysk *et al.*, 1988) and studies have specifically shown that when compared with the general population, first degree relatives of CD patients have a 10-fold increase in the risk of having the same disease as the patients (Orholm *et al.*, 1991).

A recent study published in Nature Genetics, concludes a high heritability of λs ~ 20-35 (Barrett *et al.*, 2008). A genetic predisposition by twin studies that contrast monozygotic concordance rates of 50% with only 10% in dizygotic pairs (WTCCC, 2007) has also been published. Identical twins are significantly more likely to be concordant for IBD than non-identical twins (Thompson *et al.*, 1996). In 2011, Hart & Ng published that fifteen percent of patients with CD have a relative with either CD or UC and the concordance rate in monozygotic twins is about 45% (higher than for UC). Patterns of disease within families are similar.

2.8.4 Molecular Genetic Risk Factors

There has been significant progress over the last decade in identifying susceptibility genes for CD. Approximately one-third of patients with CD have mutations in NOD2, the first CD gene identified, on chromosome 16 (Noomen *et al.*, 2009). NOD2 heterozygotes have a two-fold increased risk of developing CD compared with wild type, whereas NOD2 homozygotes have a 17-fold increased risk. NOD2 variants are particularly associated with ileal CD. NOD2 encodes an intracellular receptor for bacterial muramyl dipeptide and modulates activation of NFkB and downstream pro-inflammatory mediators by a poorly understood mechanism. GWAS studies for CD have highlighted several other important immune pathways: autophagy, a process involving the degradation of a cell's own components and

intracellular bacteria, is highlighted by association of the autophagy genes ATG16L1 (Lacher *et al.*, 2009) and IRGM with CD; and the IL-23 pathway is highlighted by association of variants in the IL-23 receptor gene (Hart & Ng, 2011). However, mutation of this gene is found in no more than 30% of people with CD, and homozygote's represent only about 5%- 10% of patients (Selby, 2003).

Many studies have reported the association between ATG16L1 gene and CD suggesting that it is an important CD susceptibility marker (Palomino-Morales *et al.*, 2009), and also, more specifically that the mode of inheritance of the G allele was most likely to be co-dominant in Caucasians (Zhang *et al.*, 2009). Another study has also identified highly significant associations of TNFSF15 (TNFα super family, member 15) from a GWAS of Japanese patients using 80,000 gene-based SNP markers. This study was also confirmed with two European IBD cohorts. Interestingly, a core TNFSF15 haplotype showing association with increased risk to the disease was common in these two ethnic groups (Yamazaki *et al.*, 2005). The OR, according to current studies, of common alleles for CD risk factors were above 1.3 or 1.5 (Barrett *et al.*, 2008).

The main determinant of disease behaviour in CD is location of disease. In addition, ileal disease has been reported to be associated with a family history and earlier onset of disease. It is therefore not surprising that NOD2 has been associated with involvement of ileal segments of the intestine. A relationship between a stricturing disease pattern and NOD2 mutations and a higher need for operations has also been suggested, but is less established. It is not clear, whether the relationship with the stenosing phenotype and surgery is a true association or reflects the high proportion of patients with ileal disease developing strictures necessitating surgical procedures. Despite the effect of NOD2 on the susceptibility and course of CD, no association

has been found between carrier ship of NOD2 mutations and response to various IBD therapies, including anti-TNF directed strategies (Noomen *et al.*, 2009).

A recent meta-analysis of GWAS identified 30 new loci for CD susceptibility including loci with functionally interesting candidate genes such as SMAD3, ERAP2, IL10, IL2RA, TYK2, FUT2, DNMT3A, DENND1B, BACH2 and TAGAP. Combined with previously confirmed loci, recent results have been able to identify 71 distinct loci (including those previously identified) with genome-wide significant evidence for association with CD (Franke *et al.*, 2010).

Based on these new results, 23% of the genetic predisposition to CD can now be explained. In line with previous results, most of these genes are involved in the control of the interactions between the intestinal bacterial flora and local immune cells in the mucosa, and in the activation of what is known as the adaptive immune response (Franke *et al.*, 2010). The new loci that have been identified in this study are interesting and noteworthy however causality data to confirm their true value was not available at the time of the study. This study implicated genes such as VAMP3, MUC1-SCAMP3, DENND1B, IL10, DNMT3A, GCKR, THADA, ERAP2, NDF1P1, CPEB4, TAGAP, IL2RA, FADS2, TNFSF11, SMAD3, TYK2, and FUT2. Many of these loci have identified potentially causal genes, though confirmation of their role must await detailed fine mapping, expression and functional studies (Franke *et al.*, 2010).

Another study published in 2011 looks at Celiac Disease (CelD) and CD and the risk loci that they collectively share (Festen *et al.*, 2011). The two diseases can co-occur within families, and studies suggest that CelD patients have a higher risk to develop CD than the general population. These observations suggest that CD and CelD may share common genetic risk loci. Two such shared loci, IL18RAP and PTPN2, have

already been identified independently for these two diseases. Nine independent regions had nominal association P-value and showed evidence of association to the individual diseases in the original scans. These include the seven loci that had not been reported as shared loci and thus were tested in additional CelD and CD cohorts. Two of these loci, TAGAP and PUS10, showed significant evidence of replication in the combined CelD and CD replication cohorts and were firmly established as shared risk loci of genome-wide significance. Through this meta-analysis of GWAS data from CD and CelD, the study identified four shared risk loci: PTPN2, IL18RAP, TAGAP, and PUS10 (Festen *et al.*, 2011).

McGovern *et al.* (2010) found supportive evidence for 21 out of 40 CD loci identified in a recent CD GWAS meta-analysis, including two loci which had only nominally achieved replication (rs4807569, 19p13; rs991804, CCL2/CCL7) that identified genetic variation in both innate and adaptive immune systems and its association with CD susceptibility. In addition, associations with genes were identified that involved in tight junctions/epithelial integrity (ASHL, ARPC1A), innate immunity (EXOC2), dendritic cell biology [CADM1 (IGSF4)], macrophage development (MMD2), TGF-beta signalling (MAP3K7IP1) and FUT2 (rs602662) ($P = 3.4 \times 10^{-5}$). The study demonstrated replication in an independent cohort between the four primary FUT2 SNPs and CD (rs602662) combined ($P = 4.9 \times 10^{-8}$), and also association with FUT2 W143X ($P = 2.6 \times 10^{-5}$). These findings strongly implicate this locus in CD susceptibility and highlight the role of the mucus layer in the development of CD (McGovern *et al.*, 2010).

As suggested before, another study has also identified three previously described mutations within the IBD susceptibility gene CARD15 (R702W, G908R, 1007fs) that increase susceptibility to CD with a terminal ileal and/or ileocolonic location and

fibrostenosing behaviour (Crawford *et al.*, 2007). The R702W mutant allele was associated with CD on case-control ($q = 0.036$, $P = 1.0 \times 10^{-4}$) analysis, and 1007fs with CD on pedigree disequilibrium testing ($P = 2.0 \times 10^{-2}$). All 3 CARD15 mutations increased susceptibility to a variety of CD sub-phenotypic manifestations, including early-onset CD in individuals with a family history of IBD, and CD complicated by extra-intestinal disease. The study also presented evidence to suggest that R702W may predispose to a more generalized form of CD. Additionally, this research confirmed that CARD15 mutations are associated with terminal ileal/ileocolonic, and to a lesser extent, also associated with fibrostenosing CD (Crawford *et al.*, 2007).

Another study identified the association of CD and sarcoidosis (SA) in which 24 SNPs that were the most strongly associated in the combined CD/SA phenotype were selected for verification in an independent sample of patients (Franke *et al.*, 2008). The most significantly associated SNP was rs1398024 on chromosome 10p12.2. During further testing, a significantly moderate association was observed between rs1398024 and UC, but not CD. Extensive fine mapping of the 10p12.2 locus points to yet unidentified variants in the C10ORF67 gene region as the most likely underlying risk factors. This study demonstrated that the combined analysis of different, albeit clinically related, phenotypes can lead to the identification of common susceptibility loci (Franke *et al.*, 2008).

Several risk factors for CD were also identified in another recent GWAS which combined data from three studies on CD (a total of 3, 230 cases and 4, 829 controls) and was analysed (Barett *et al.*, 2008). Replication was carried out in 3,664 independent cases with a mixture of population-based and family-based controls. While the individual scans did identify new risk factors, they were only well-

powered to discover common alleles with ORs above 1.3 (in the case of the WTCCC) or 1.5 (Barett *et al.*, 2008). By contrast, the combined sample had 74% power at an OR of 1.2, allowing evaluation of the role of alleles with smaller effect sizes for the first time. The results strongly confirmed 11 previously reported loci and provide genome-wide significant evidence for 21 additional loci, including the regions containing STAT3, JAK2, ICOSLG, CDKAL1 and ITLN1 (Barett *et al.*, 2008). Previously unrecognized pathogenic mechanisms of IBD have also been acknowledged in this study, including the importance of autophagy and innate immunity in the behaviour of CD, as well as also highlighting genetic associations between CD and other auto-inflammatory conditions (Barrett *et al.*, 2008).

A GWAS in individuals with CD by the WTCCC detected strong association at four novel loci in 2007 (Parkes *et al.*). The study tested 37 SNPs from these and other loci for association in an independent case-control sample. Replication for the autophagy-inducing IRGM gene on chromosome 5q33.1 and for nine other loci *(replication* $P=6.6 \times 10^{-4}$*, combined* $P = 2.1 \times 10^{-10}$), including NKX2-3, PTPN2 and gene deserts on chromosomes 1q and 5p13 were obtained (Parkes *et al.*, 2007).

A multi-stage genome-wide scan of 393 German CD cases and 399 controls tested and confirmed 161 SNPs, an association with the known CD susceptibility gene NOD2, the 5q31 haplotype, and the reported CD locus at 5p13.1 (Franke *et al.*, 2007). In addition, SNP rs1793004 in the gene encoding Nell-like 1 precursor (NELL1, chromosome 11p15.1) showed a consistent disease-association in independent German populations and family-based samples (Franke *et al.*, 2007). Subsequent fine mapping and replication in an independent sample supported the authenticity of the NELL1 association and indicated that NELL1 is a ubiquitous IBD susceptibility locus. The novel 5p13.1 locus was also replicated in the

French/Canadian sample and in an independent UK CD patient panel (453 cases, 521 controls) (Franke *et al.*, 2007). Several associations were replicated in at least one independent sample, and point to an involvement of ITGB6 (upstream), GRM8 (downstream), OR5V1 (downstream), PPP3R2 (downstream), NM_152575 (upstream) and HNF4G (intron) (Franke *et al.*, 2007).

Rioux *et al.* (2007) also identified several new regions of association to CD, specifically, in addition to the previously established CARD15 and IL23R associations, strong and significantly replicated associations with an intergenic region on 10q21.1 and a coding variant in ATG16L1 was also identified. This study also reported strong associations with independent replication to variation in the genomic regions encoding PHOX2B, NCF4 and a predicted gene on 16q24.1 (FAM92B) (Rioux *et al.*, 2007). It was demonstrated that ATG16L1 is expressed in intestinal epithelial cell lines and that functional knockdown of this gene abrogates autophagy of *Salmonella typhimurium*. Together, these findings suggest that autophagy and host cell responses to intracellular microbes are involved in the pathogenesis of CD (Rioux *et al.*, 2007).

There are human leukocyte antigen (HLA) class II genes that have been linked with CD as well, the DR1/DQw5 and DRB3*301 haplotypes (Stenson, 2003). Previous, linkage analysis has acknowledged two susceptibility loci; IBD1 on chromosome 16 and IBD12 on chromosome 12 (Stenson, 2003). In addition, the recognition of a relationship between CD and the pericentromeric region of chromosome 16 (IBD1) by Hugot in 1996, generated a series of genomic scans and linkage analyses in exploration of susceptibility and phenotypic modifier genes (Mamula *et al.*, 2003).

In 2001, the discovery that particular polymorphisms in the CARD15/NOD2 gene at the IBD1 locus were associated with CD stimulated a new area of genotype-phenotype research at that time (Mamula *et al.,* 2003). Although NOD2 mutations are seen much more commonly in CD than in the general population, only a small portion (~15%) of all CD patients have a mutation in NOD2, demonstrating that CD is genetically heterogeneous and suggests that other mutations in other genes may contribute to CD in other populations (Stenson, 2003).

Another approach to understand the genetic basis of CD is the search for subclinical markers i.e. parameters used to detect abnormal genotype (Stenson, 2003). The subclinical markers that have received the most attention are increased intestinal epithelial permeability in CD (Stenson, 2003). There is a report of increased intestinal permeability to PEG-400 in CD patients and their first degree relatives (Stenson, 2003).

In Australia, a study was carried out in 2003 (Cavanaugh *et al.*) that implicated three risk NOD2 alleles for CD within the Australian population: (Arg702Trp (C/T), Gly908Arg (G/C) and 980fs981 (-/C). This study demonstrated that the three risk alleles are more frequent in CD, than in controls, with allelic frequencies of 0.11, 0.02 and 0.07, respectively. Heterozygosity for individual variants conferred a three-fold increase in risk for CD while substantially higher risks were associated with being homozygous or compound heterozygous. Despite a significantly lower population allele frequency for the frame-shift mutation than reported by other groups, this study showed a similar contribution by this allele to the risk of developing CD. The investigators of this study concluded that while the three risk alleles influence susceptibility to CD in Australia, it showed that these alleles do not fully explain the linkage evidence and suggest that there are likely additional IBD1

susceptibility alleles yet to be described in Australian CD at the NOD2 locus. The study also show a second linkage peak in Australian CD that provides some support for a second disease susceptibility locus on chromosome 16.

These and many other evidences suggest that IBDs, including CD, occur as a result of an inappropriate inflammatory response in which genetic and environmental factors play important roles. Identifying SNPs in individuals can improve differential diagnosis and optimize treatment efficacy. Much remains to be learned about the prevalence, influence, and interactions between specific genetic polymorphisms and human IBDs such as CD. However, the future understanding of these factors combined with the ability to identify these SNPs easily in individuals might impact the potential to improve differential diagnosis and optimize treatment efficacy.

CHAPTER THREE: KEY CONCEPTS AND METHODOLOGIES

3.1 Key Concepts

The scope of this project involves concepts from a variety of scientific and/or technical backgrounds; e.g. the knowledge of clinical and/or medical concepts, and familiarity with bioinformatics concepts as well as genetics concepts. Brief overviews of some of the key concepts that have been utilised in this research project are provided below so that the understanding of the information provided in subsequent chapters is comprehendible for an individual from any background. Specifically, it is hoped that the results that have been achieved from this research can be understood precisely with the brief clarification of some key concepts of this study that have been provided. However, this description is by no means a thorough explanation of these key concepts; it is only a very concise outline.

3.2 Genetics

The science of genetics is a discipline of biology which explores genes, heredity and variation in living organisms. The subject of genetics concentrates on the molecular structure and function of genes. Gene behaviour in the context of a living cell or organism, studies of patterns of inheritance, the distribution of genes, gene variations and change in populations are all part of the science of genetics. The role of genetics in the development, behaviour and appearance of living organisms is predetermined, and it is with the combination of other external experiences that determines the eventual outcomes that manifest in any living organism. Some of the basic concepts that are incorporated in this project that revolve around the topic of genetics are further described in brief subsequently.

3.2.1 Deoxyribonucleic Acid

Deoxyribonucleic Acid (DNA) is a molecule of information encoding the genetic instructions that are used in the development and functioning of all known living organisms; it is one of the known macromolecules that are essential for all known forms of life. It is a self-replicating material which is existent in almost all living organisms as the key constituent of chromosomes. A set of chromosomes in a cell creates up its genome; the human genome has nearly 3 billion base pairs of DNA organized into 46 chromosomes (Venter *et al.*, 2001).

The information carried on DNA is held in a sequence of fragments called genes. A gene is a region of DNA that influences a specific characteristic in an organism. Only a minor fraction of the DNA consists of genes (about 1%). Genes are sequences of bases that define a particular sequence of amino acids. There are 20 different amino acids that can be used in sequences. Amino acids join together in the prescribed sequence by a structure called a ribosome, then folded and packed. This packed amino acid sequence is called a protein.

Only a specific portion of DNA within a gene is converted (translated) into amino acids. A complete set of the genetic data held within genes on a genome is called an organism's genotype. Genetic information is carried in each molecule of DNA which consists of two strands that are coiled round each other to form a double helix, a structure like a spiral ladder. Every rung of the ladder consists of a pair of chemical groups called nucleotides. There are four types of nucleotides (or bases), namely adenine (A), cytosine (C), guanine (G) and thymine (T). These bases combine in specific pairs (A-T & C-G) so that the sequence on one strand of the double helix is complimentary to that on the other i.e. they are anti-parallel. It is the specific

sequence of these nucleotides which establishes the genetic information of an individual. Transmission of genetic information in genes is achieved by complementary base-pairing. The DNA double helix is stabilised by hydrogen bond forces between nucleotides. Within cells, DNA is organized into long structures called chromosomes. During cell division, these chromosomes are duplicated in the process of DNA replication, providing each cell its own complete set of chromosomes. The expression of genes is influenced by how the DNA is packaged in the chromosomes.

Nucleotide modifications during DNA packaging can cause regions of low or no gene expression. DNA can also be damaged by a range of mutagens that change the sequence of the DNA, e.g. X-rays, electromagnetic radiation, ultraviolet light and oxidizing agents. Various forms of DNA damage to nucleotides can take place e.g. insertions (of nucleotides), deletion (of nucleotides), point mutations (single nucleotide substitution) and chromosomal translocations. Most of the normal cellular processing damage incurred by DNA is repaired using DNA repair mechanisms, however DNA damage caused by mutagens may not all be reparable.

3.2.2 Mutations

Mutations are the changing configuration of a gene that results in a variant form triggered by the modification of single base units (nucleotides) in DNA. Mutations are generated as a result of unrepaired damage to DNA, usually caused by mutagens, errors in DNA replication processes, or the deletion, insertion, or rearrangement of larger sections of genes or chromosomes. Each DNA variant is known as an allele; common variants are referred to as major alleles, and rare variants are referred to as

minor alleles. Mutations do not automatically always cause a change in the subsequent characteristic (phenotype) of an organism.

Mutations are the main cause for evolution, cancer and other biological processes. The changes to DNA caused by mutations can alter the product of a gene, prevent the gene from functioning appropriately or cause changes in proteins produced by the gene. Organisms have developed DNA repair mechanisms to deal with DNA mutations. However, these repair mechanisms do not always work. In those circumstances where these repair mechanisms fail to rectify damage caused by mutations, changes to organism phenotype will develop. Phenotype changes in an organism can be both beneficial (evolution) and detrimental (disease).

Typically mutations that alter protein sequences can be harmful to an organism, although on occasion of a specific environment; the effect may be able to produce positive effects. Detrimental mutations, which occur in most occasions, produce partially or completely non-functional proteins. Each cell, in order to function properly, depends on thousands of proteins to function in the correct places at the correct times. When a mutation alters a protein that plays a critical role in a human, the subsequent effect is typically a medical disorder. A genetic condition can occur in humans with mutations in one or more genes. If mutations occur in germ cells, than offspring will carry that mutation in all of its cells; as is the case of hereditary diseases. As a result, identifying and examining the underlying genetic contribution to disease is necessary to understand the pathology of a disease.

3.2.3 Single Nucleotide Polymorphisms (SNPs)

A SNP is a variation in the DNA sequence that occurs as a result of a change in a single base of DNA. Almost all common SNPs have only two alleles i.e. one of the

two alternative forms of the same gene or same genetic locus (group of genes). A number of factors determine allele selection including natural selection and other factors like genetic recombinations. Within a population, SNPs can be assigned a minor allele frequency - the lowest allele frequency at a locus that is observed in a particular population. This is simply the smaller of the two allele frequencies for SNPs. There are dissimilarities between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in a different ethnic group. Genetic variations underlie differences in our susceptibility to disease. Genetic variations influence the severity of disease we encounter and how our body reacts to treatments. Variations in the DNA sequences of humans can affect how humans develop diseases. SNPs can be used for comparing regions of the genome between cohorts in GWAS. The study of SNPs is also critical for personalized medicine techniques. Bioinformatics databases for SNPs (e.g. dbSNP from the National Centre for Biotechnology Information (NCBI)) exist that can be used for such types of analysis.

3.3 Genomics and Bioinformatics

Genomics is the branch of molecular biology that is concerned with the structure, function, evolution and mapping of genomes. It is a discipline in genetics that applies various techniques such as DNA sequencing and bioinformatics to sequence, assemble and analyse the function and structure of genomes. This field of study also includes the study of entire DNA sequences of organisms and fine-scale genetic mapping. It also includes studies of intragenomic phenomena such as epistasis, and other interactions between loci and alleles within the genome.

In order to understand and analyse a genome, different sequencing mechanisms have evolved over time. Historically, sequencing was done in centralized facilities containing research laboratories with costly instrumentation and necessary technical support. As sequencing technology continues to improve, a new generation of effective fast turnaround bench top sequencers have become readily obtainable for the average academic laboratory. Overall, genome sequencing approaches are divided into two main methods: shotgun and high-throughput sequencing. Shotgun sequencing is a method designed for the analysis of DNA sequences longer than 1,000 base-pairs, up to and including entire chromosomes. The high demand for low cost sequencing has driven the development of high-throughput sequencing, also known as next-generation sequencing (NGS). These NGS method technologies allow the production of thousands of millions of sequences at once. NGS technologies are able to lower the cost of DNA sequencing beyond what was originally possible.

3.3.1 Genomic Technologies

Genomic technologies are best defined as those technologies that are used to analyse and interpret genomic information. The recent history of genomics has been driven by numerous technological advancements. The primary advancement at the very beginning was the methodologies that were developed of the polymerase chain reaction (PCR) and automated DNA sequencing. PCR methods allowed the amplification of usable amounts of DNA from very small amounts of starting material. Previous methods that allowed DNA sequencing of molecules one by one are being substituted by methods where billions of DNA molecules are sequenced simultaneously (Conley *et al*, 2013). Automated DNA sequencing methods have evolved to the point that the entire DNA sequence of microbial genomes comprising

of several million base pairs can be acquired in relatively small amounts of time (Baldi and Hatfield, 2002).

The evolution of the power of genomic analysis techniques began with the invention of DNA cloning in the 1970's. The historical impact of these technologies is clearly immense. While the origins of many technical progresses were rooted in the pre-cloning era, most of the technology derives from the last quarter of the 20th century. The revolutionary technology of DNA cloning opened the world of macromolecular information in the genomes of living things, and biologists began to clone and study the structure and function of individual genes (Galas and McCormack, 2003).

Since the first genome was completely sequenced in 1977; the sequencing of whole genomes as well as of individual regions and genes has become a major focus of modern biology and completely transformed the field of genetics. Initially, DNA sequencing was a barely automated and very tedious process which involved determining only a few hundred nucleotides at a time. Gradually, in the late 1980's, semi-automated sequencers with high throughput became available, but were still only able to determine a few sequences at a time. In the early 1990's, a breakthrough lead to the development of capillary array electrophoresis and appropriate detection systems (Kircher and Kelso, 2010). From 1990 to 1993 the positional cloning process underwent transformation, yielding new technologies for the human genome project. In 1991, Craig Venter at the National Institutes of Health (USA) developed a way of finding human genes that did not require sequencing of the entire human genome (Baldi and Hatfield, 2002). This key technology conferred the ability to clone and characterise successively larger pieces of human genomic DNA. The positional cloning of human genes now became a real technical possibility as a routine systematic process (Galas and McCormack, 2003). In 1996, developments

advanced to the production of a commercial single capillary sequencer (ABI prism 310). Then, in 1998, the GE Healthcare MegaBACE 1000 and the ABI Prism 3700 DNA Analyser became the first commercial 96 capillary sequencers. This development was labelled as high-throughput sequencing (Kircher and Kelso, 2010).

Over the last decade, alternative sequencing strategies have become available which have completely redefined high-throughput sequencing technologies. These developments out-perform the older sequencing technologies by a factor of 100 to 1,000 in daily throughput, and at the same time reduce the cost of sequencing one million nucleotides (1Mb) to 4-0.1% of that associated with older sequencing (Kircher and Kelso, 2010). To reflect these advancements and developments, several companies and researchers have termed this the "next-generation sequencing" (NGS). NGS sequencing have dramatically increased the output, reduced the time and cost of sequencing, and allowed for greater coverage of the genome. NGS technology has made it possible to sequence the human genome in a matter of days, whereas the first human genome was sequenced after thirteen years of international effort and at a cost of nearly $3 billion dollars (Conley *et al*, 2013).

Two major projects that utilised this field of study are the Human Genome Project (BERIS, 2011) and the HapMap (The International HapMap Consortium, 2013). The Human Genome Project was a research project that aimed to generate a human genome reference sequence in about ten years. The project was launched in 1990 and the first draft human genome sequence was completed in late 2000. The sequence was essentially complete in April 2003, although some regions still remain (particularly long lengths of repeated DNA) that are very difficult to unambiguously determine the sequence for using current technology. Between July 2000 and October 2009, there have been 18 revisions of the reference sequence, and this data

is publicly available on the US National Institute of Health (NIH) website. The goal of the International HapMap Project is to develop a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation. The HapMap is expected to be a key resource for researchers to use to find genes affecting health, disease, and responses to drugs and environmental factors. To develop the HapMap, the samples will be genotyped for at least 1 million SNPs across the human genome. When the HapMap project started, 2.8 million SNPs were in the public database called dbSNP. However, many chromosome regions had too few SNPs, and many SNPs were too rare to be useful, so millions of additional SNPs were needed to develop the HapMap. The project discovered another 2.8 million SNPs by September of 2003, and since then SNP discovery continues. In 2004, the project initially produced a map of 600,000 SNPs evenly spaced across the genome, which is a density of one SNP every 5000 bases. Additional SNPs will be genotyped where needed to define haplotypes.

With the current abundance of massive biological datasets, computational studies have become one of the most important means to biological discovery. Different approaches to study large-scale genomic datasets have been developed, including the GWAS approach which is explained in the next section. The improvements in genomic data collection technologies have been an important driving force behind the increased accuracy and rapidity in which we can currently collect genomic data. Improved accuracy and rapidity of data collection are also very important aspects of moving genomic findings into the clinical arena for translation to patient care in the near future.

3.3.2 Computational Genomics

Computational genomics refers to the use of computational analysis mechanisms to decipher biological information from genome sequences and related data (Koonin *et al.*, 2001). It is considered to be a subset of bioinformatics, focusing on the use of whole genomes (rather than individual genes) to understand DNA principles of a species, and how the DNA controls its biology at the molecular level. The recent advancements of NGS technology is allowing researchers to inexpensively generate a large volume of genome sequence data; the challenge now lies in developing computational genomic methods for integrating different data types and extracting complex patterns accurately and efficiently from a large volume of data.

In the context of genomic sequence data, various computational methods have been developed that have been able to perform gene delineation or gene-finding tasks. These methods of automated annotation are critical to making full use of the exponentially increasing volume of genomic sequences being generated and released in databases for research and commercial purposes. One common and useful ordering of computational genomics approaches follows the pipeline of large-scale genome examinations performed in various situations. These techniques arise from the 'mainstream' bioinformatics undertakings and mostly concentrate on genome sequence analysis, such as gene finding, sequence diagnostics, database searching, sequence clustering and functional annotation.

Baldi and Brunak (2001) critically explain:

> *"As genome and other sequencing projects continue to advance unabated, the emphasis progressively switches from the accumulation of data to its interpretation. The large amounts of data (collected) create a critical need for*

theoretical, algorithmic and software advances in storing, retrieving, networking, processing, analysing, navigating and visualizing biological information. Large databases of biological information create both challenging data-mining problems and opportunities, each requiring new ideas. In this regard, conventional computer science algorithms have been useful, but are increasingly unable to address many of the most interesting sequence analysis problems. This is due to the inherent complexity of biological systems, brought about by evolutionary tinkering, and to our lack of comprehensive theory of life's organization at the molecular level".

Traditionally, the one-SNP-at-a-time approach to analysing large-scale genomic data sets relied on SNP based association data, i.e. choosing the most statistically significant single SNP. This has been able to provide relatively little information in terms of overall disease risk for any individual. GWAS have linked genetic variants to probabilities of disease associations and report associations for individual SNPs. However, functional information typically exists for proteins or genes and therefore SNPs have to be confined to genes and their individual association signals combined. Since GWASs are currently moving from microarray based technology towards NGS methods, NGS in principle, allows for the identification of all genetic variants.

Nevertheless, the information that is collected from GWAS can be still be utilised with additional analysis provided by the application of computational genomic methods. Currently evolving computational genomics methods are now becoming rapidly more focused on analysing multiple SNPs at a time, to the consideration of entire pathways comprising of possibly dozens of genes and their environmental interactions, and even analysing multiple pathways that interrelate with each other in complex systemic networks. With the advent of evolving genomic tools, not only has

the characterization of the full spectrum of genetic variations within candidate genes become possible, but genome-wide analysis searching for gene associations and interactions are increasingly becoming more powerful and complex.

Machine learning methods that rely on bioinformatics knowledge are essentially important since the amount of biological data requiring automated analysis has exponentially exploded. Machine learning methods provide the technical basis of data mining and are used to extract information from raw data that presides in large-scale genomic databases. Other methods include methods that have been created to determine SNP-SNP interactions and SNP-environment interactions in order tolocate associations with complex diseases. These methods rely on large-scale data mining techniques that are utilised by the application of computational and statistical algorithms. Identifying SNP-SNP, gene-gene, or SNP-environment interactions are biologically convincing models for disease genetics and can provide many advantageous insights on the complex mechanisms and pathways that are used in the creation of complex diseases.

Using CD data many studies have experimented with a variety of computational statistical methods that investigate CD genetic interactions. An example has been illustrated by Dinu *et al* (2012), where SNP-SNP interactions in CD genetics using logic regression were investigated. This study specifically considered two specific forms of biologically plausible SNP-SNP interactions, 'SNP-intersection' and 'SNP Union' and analysed the CD GWAS data of the WTCCC using a limited form of logic regression. This study found strong evidence of CD-association for 195 genes, (e.g., *ISX*, *SLCO6A1*, *TMEM183A*) as well as confirmed many previously identified susceptibility genes in CD GWAS (e.g., *IL23R*, *NOD2*, *CYLD*, *NKX2-3*, *IL12RB2*, *ATG16L1*). Notably, 37 of the 59 chromosomal locations indicated for CD-

association by a meta-analysis of CD GWAS, involving over 22,000 cases and 29,000 controls, represented in 195 genes, as well as some chromosomal locations previously indicated only in linkage studies but not in GWAS, were identified using their methodological approach. The analysis was repeated with two smaller GWASs from the dbGaP and in spite of differences of populations and study power across the three datasets, consistencies were observed across the three datasets. Notable examples included *TMEM183A* and *SLCO6A1* which exhibited strong evidence consistently in the WTCCC and both of the dbGaP SNP-SNP interaction analyses using the methods developed as part of this study (Dinu *et al*, 2012).

Another study performed by Okazaki *et al* (2008) uses a multi-dimensionality reduction method approach to explore gene-gene interactions of CD. The methods developed as part of this study identified ATG16L1, IBD5, and IL23R SNPs as significantly associated with CD. A multivariate analysis showed independent CD association for carriers of IBD5, IGR2230 and IL23R- rs10889677 while retaining association for NOD2 mutation carriers, IBD family history, tobacco, and Jewish ethnicity. IL23R minor variants for Arg381Gln and Intron 6- rs7517848 showed independent, CD protection and 3' untranslated variant rs108896778 showed risk. The MDR analysis suggested an interaction between IBD5, ATG16L1, and IL23R risk alleles. This study was a population-based analysis of CD risk factors useful for characterizing the epidemiology of multiple CD genetic and non-genetic risk factors and the study suggested that gene-gene interactions were likely but required further evaluation in larger population based cohorts (Okazaki *et al*, 2008).

In the same way, another study investigated a case-only design method by performing logistic regression analyses to investigate statistical interactions between NOD2 risk alleles and smoking status (Helbig *et al*, 2012). This study detected a

significant negative interaction between carriership of at least one of the NOD2 risk alleles and history of ever having smoked as well as smoking at the time of CD diagnosis, as a result of the methods applied. The study observed a significant negative gene-environment interaction which suggests that the risk of increase for CD conferred simultaneously by cigarette smoking and that the 1007fs NOD2 polymorphism is smaller than expected and may point to a biological interaction. This investigation carried out as a result of the statistical methods developed, suggests a confirmation of a mechanistic interaction by which cigarette smoking modulates the risk associated with disease-associated NOD2 alleles which may be important in not only better understanding the biological mechanisms underlying CD pathogenesis, but also in developing appropriate recommendations for individuals at high risk of developing CD (Helbig *et al*, 2012).

As more research is performed, technologically advanced and more powerful computational methods will continue to be developed. This research project also adopts a number of step-wise data reduction methods in order to identify CD associated risk factors. As part of this research, GWAS data obtained from the WTCCC and dbGaP were analysed using these step-wise data reduction methods and are explained in subsequent sections of this thesis.

3.3.3 Genome Wide Association Studies

The methods of GWAS, also known as whole genome association studies (WGAS) are large scale investigations of many common genetic variants in different individuals to identify variants associated with a specific disease trait. This approach of analysis looks at polymorphisms or genetic markers that display a relationship with a trait of interest. The GWAS approach is to investigate the entire genome at a

time. GWAS rely on associations between SNPs and traits for major diseases. These studies compare the DNA of two groups of participants i.e. people with the disease (cases) and demographically similar people without the disease (controls). If one type of the variant is more frequent in cases, the SNP is identified as associated with the disease. The associated SNPs of the disease are then located on a region of the human genome and this region is then marked as influencing the risk of disease. Because GWAS identify SNPs in DNA associated with a disease, they cannot specify causal genes.

Data for genetic markers throughout the genome can be collected using a variety of technologies, including those that focus on specifically selected polymorphisms throughout the genome (such as microarray technologies) and those that collect data for all variants throughout the genome (such as NGS).

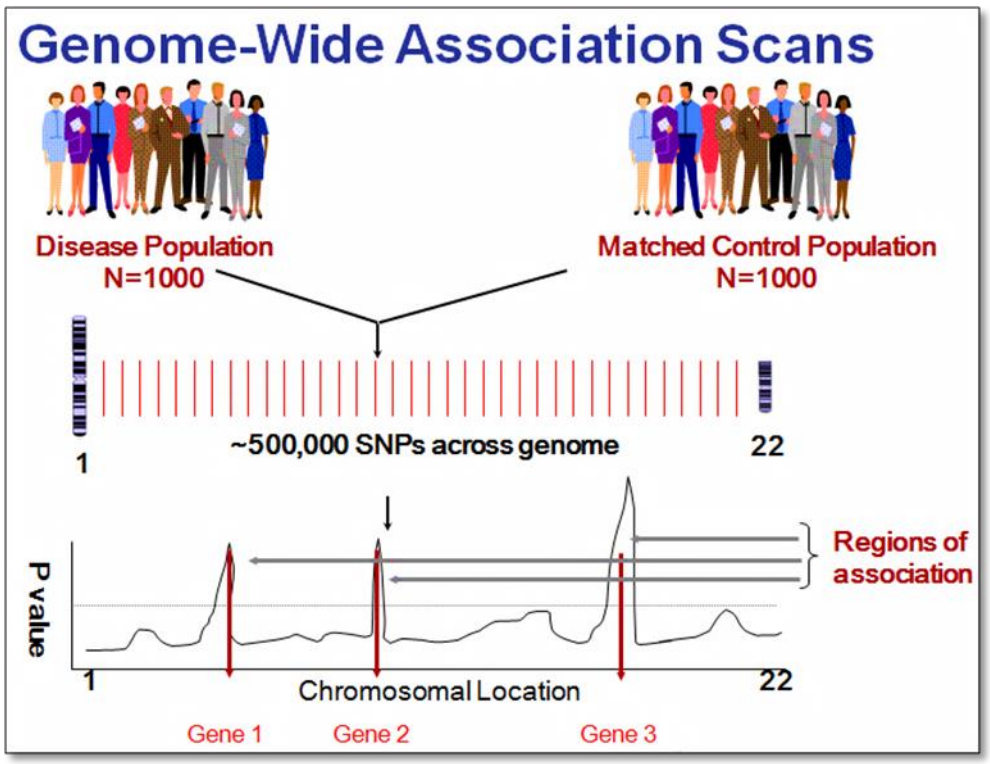

Figure 3.1 Genome Wide Association Scan Description (Lea *et al*, 2009)

Recently, studies have indicated that hundreds or thousands of individuals have been tested, over 1,200 human GWAS have been examined, over 200 diseases and traits, and almost 4,000 SNP associations have been identified (Johnson, 2009). The idea behind GWAS is to study and compare cases with controls from two large groups of individuals affected by a disease. All individuals in each group are genotyped for common known SNPs. Significantly different allele frequencies of each SNP are then investigated between the case and the control group. In such studies, effect sizes are measured using the ORs which report the ratio between two proportions, which in the context of GWAS are the proportion of individuals in the case group carrying a specific allele, and the proportion of individuals in the control group with the same allele. If the allele frequency in the case group is much higher than the control group, the OR will also be higher than a value of one (1). A P-value for the significance of the OR is calculated using a simple Chi-Squared test. The aim of the GWAS approach is to find ORs that are significant and greater than one, showing that there is SNP association with disease.

The goal and challenge of the GWAS approach is to apply results obtained from GWAS into clinical applications, accelerating drug and diagnostics development. Identifying risk-SNP markers as a means to directly improve prognosis accuracy is a critical role of GWAS. GWAS have a number of limitations as well, common problems being the lack of well-defined case and control groups, small sample sizes, multiple testing controls and control for population stratification. These limitations can be dealt with through proper quality control and study setup methods.

Despite the realisation of many perceived failures of GWAS since the first large scale GWAS study by WTCCC in 2007, researchers must remember that the main contribution of GWAS was always going to be obtaining biological insights from its

discoveries, a contribution that will remain relevant if many genes remain unidentified (Visccher *et al.*, 2012). The GWAS approach has actually captured a substantially greater proportion of the genetic risk than has previously been recognised. Visccher *et al.* (2012) clearly summarises the scope of GWAS and points out that:

> *"It is surprising how right from the start critics have taken such a dogmatic view and have not put subsequent research findings into perspective. In a period of only five years, a fantastic treasure trove of new, biologically relevant discoveries has been generated and researchers can be confident that this will continue for the foreseeable future, in particular when combined with the opportunities arising from the new genomics technologies. It's interesting to observe that in historical and ongoing criticism of GWAS the goalposts keep moving. First it was 'it won't work' then 'only a few genes are identified for any particular disease', then 'there is no biological relevance', followed by 'there is no clinical relevance' and finally 'the results are not yet translated into clinical practice'. Clearly, an objective view of this particular experimental design is that it has been very successful in a very short period of time. For the future, the technological advance to sequence entire genomes in large samples at affordable prices is likely to generate additional genes, pathways and biological insights, as well as the identification of causal mutations*."

Important and meaningful results that have been gained as a result of GWAS include the discovery of the autophagy pathway in CD. Applications of GWAS using human populations have allowed investigators to identify significant and new knowledge regarding the involvement of genes and pathways associated with common complex

diseases. An immense abundance of novel biological insights have been discovered and have allowed scientists to perform further research into incorporating these discoveries into direct procedures for clinical practice.

Since the GWAS, the concept of Whole Genome Sequencing (WGS) has also emerged, enabling the sequencing of entire genomes in large samples at affordable prices and this technique is likely to generate additional genes, pathways, and biological insights, as well as to identify causal mutations. Only recently have GWA studies begun to emerge as a means to identify associated disease risk factors. Following the identification of several disease-associated polymorphisms by GWAS, interest is now focusing on the detection of effects that, owing to their interaction with other genetic or environmental factors, might not be identified by using standard single-locus tests. In addition to increasing the power to detect associations, it is hoped that detecting interactions between loci will allow researchers to elucidate the biological and biochemical pathways that underpin disease (Cordell, 2009).

With the recent success of GWAS and WGAS in their ability to identify significant susceptibility genes for many common complex diseases, some scientists have also availed epidemiologic study opportunities to identify susceptible environmental risk factors of disease. Still, research into gene-environment interactions remains quite small, and will eventually begin to accelerate as investigators integrate analysis of genome-wide variation and environmental factors (Khoury & Wacholder, 2008). Since the availability of relatively cheap GWAS and companies promoting personalised genomic scans, a scientific era that has begun to realize the importance of non-genetic factors in disease risk analysis is slowly developing. We have known for decades that failure to incorporate both genetic and environmental factors in a

joint analysis will weaken the observed associations between a true risk factor and disease occurrence (Khoury & Wacholder, 2008).

Theoretically, if we are able to measure gene-environment interactions, we should sharpen our measurements of effects in subsets of the population and even potentially increase our statistical power in measuring such effects (Khoury & Wacholder, 2008). Due to the many underlying challenges that analysing genome-environment interactions causes, gaining accuracy realistically instead of just theoretically still requires time. The development of new methods to analyse Genome-environment-wide interaction studies (GEWIS) of disease occurrence in human populations is an example of methodological process leading to practical advantages (Chatterjee & Wacholder, 2009). Many approaches are being tackled, and this research project also attempts to create a model whereby genetic, environmental and clinical factors can be investigated in combination to identify an overall disease risk prediction for a more accurate diagnosis.

The traditional GWAS design and analysis that has been explained thoroughly in this section still has many drawbacks, despite the relatively significant information that has been gathered from GWA studies. The various issues and limitations that arise from GWA studies can usually be taken care of through proper quality control and study setup, however the need for more powerful and intelligent computational approaches to analysing human genomic information that can eventually be translated into clinically relevant and useful information still exists. Lack of well-defined case and control groups, insufficient sample sizes, control for multiple testing and control for population stratification are common problems that arise as a result of GWA analysis (Klein *et al*, 2012).

The most significant limitation is that the analysis of one SNP at a time poses significant limits of effect size on complex multi-factorial diseases such as CD. Klein *et al* (2012) states that risk prediction for an individual usually cannot be derived even from large-scale GWAS data and sample size is not a quality marker of GWAS per se, especially in terms of clinical relevance. When it comes to potential personalized medicine, effect size appears to be clearly the most important aspect of GWAS (Klein *et al*, 2012).It has been noted that the GWAS approach can be quite problematic because the massive number of statistical tests performed presented an unprecedented potential for false-positive results (Pearson and Manolio, 2008) and ignoring these often correctible issues has been cited as contributing to a general sense of problems with the overall GWAS methodology (Pickrell *et al*, 2011).

As Klein *et al* (2012) strategically concludes, medical science is still far from the GWAS-based personalized medicine promised in 2007. Nonetheless it is important to consider that GWAS are based on common variants that are frequently in linkage disequilibrium with the actual causative variant, which may be associated with larger effect sizes than the common variant included in the GWAS. This provides important consideration to fuel legitimate hope that genetics will continue to become an integral part of a modern medicine more specifically tailored to individual patients.

3.3.4 The Wellcome Trust Case Control Consortium

The WTCCC was created after the introduction of the SNP chip technology was introduced to the world. This technology allowed relatively cheap genome-wide genotyping and consequently the development of association studies that covered the genome as a whole. The goal of the WTCCC was to identify novel genetic variants associated with common complex diseases (WTCCC, 2007). Seven common

diseases (bipolar disorder, coronary artery disease, CD, hypertension, rheumatoid arthritis, Type I Diabetes and Type II Diabetes) were included in the initial study for testing associations. This study successfully identified many new genetic associations of the seven diseases that were investigated.

Upon application to the WTCCC data access committee, individual genotype data for any of the seven diseases is available for research purposes of disease susceptibility via genetic variations. This data set allows researchers with access to a large scale collection of data that has the potential of being quite useful for further research initiatives. A second phase of the WTCCC, the WTCCC2, has also now been established through the cooperation of collaborators after the success of the WTCCC was witnessed. The WTCCC2 aims to type over 60,000 participants at over 600,000 genome-wide locations for 13 or more common disease conditions. Data collection for the WTCCC2 is still ongoing, and it is projected that further novel insights into genetic backgrounds of these common complex diseases will be recognised.

3.3.5 The database of Genotypes and Phenotypes

The database of Genotypes and Phenotypes (dbGaP) was developed by the National Centre for Biotechnology Information (NCBI) (U.S. National Library of Medicine), in order to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include GWA studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits.

The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided researchers with powerful tools that eventually allow for the generation of the massive amount of genotypic data required to make these analyses

possible. The dbGaP provides two levels of access: open and controlled, in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization (dbGaP, 2012)

3.4 Statistical Concepts

Performing statistical analyses is a fundamental concept for any research in order to verify the relevance and/or significance of any results that are derived from the investigation. Likewise, this research project involved a number of statistical tests that were used throughout the study. These tests are briefly explained next.

3.4.1 Association Statistics

Genetic association studies are used to find candidate genes or genome regions that contribute to a specific disease by testing for a correlation between disease status and genetic variation (Lewis *et al.*, 2009). Genetic association studies test for a correlation between disease status and genetic variation to identify candidate genes or genome regions that contribute to a specific disease. A higher frequency of a SNP allele or genotype in a series of individuals affected with a disease can be interpreted as meaning that the tested variant increases the risk of a specific disease. SNPs are the most widely tested markers in association studies, but microsatellite markers, insertion/deletions, variable-number tandem repeats (VNTRs), and copy-number variants (CNVs) are also used.

Association studies are a major tool for identifying genes conferring susceptibility to complex disorders. These traits and diseases are termed "complex" because both genetic and environmental factors contribute to the susceptibility risk. Extensive experience in genetic studies for many complex disorders (such as diabetes, heart disease, autoimmune diseases, and psychiatric traits) confirms that many different genetic variants control disease risk, with each variant having only a subtle effect.

Associations with polymorphisms in candidate genes have been confirmed in many different diseases (Lohmueller *et al*., 2003), and GWA studies are identifying many novel associations in genes that had not been strong a priori candidates for the disease under test (WTCCC, 2007). However, the modest increase in risk implies that large well-designed and analysed studies are required to detect and confirm signals for association (Lewis *et al.*, 2009).

In order to calculate and identify associations, the Chi-Squared test can be used as part of the overall analysis. A Chi-Squared test, also referred to as Chi-Square test or $\varkappa^2$ test, is any statistical hypothesis test in which the sampling distribution of the test statistic is a Chi-Squared distribution when the null hypothesis is true, or any in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a Chi-Squared distribution as closely as desired by making the sample size large enough.

If a sample of size *n* is taken from a population having a normal distribution, then there is a well-known result which allows a test to be made of whether the variance of the population has a pre-determined value. For example, a manufacturing process might have been in stable condition for a long period, allowing a value for the variance to be determined essentially without error. Suppose that a variant of the

process is being tested, giving rise to a small sample of product items whose variation is to be tested. The test statistic, *T*, in this instance, could be set to be the sum of squares about the sample mean, divided by the nominal value for the variance (i.e. the value to be tested as holding). Then *T* has a Chi-Squared distribution with $n-1$ degrees of freedom. For example if the sample size is 21, the acceptance region for T for a significance level of 5% is the interval 9.59 to 34.17 (Corder *et al.*, 2009)

3.4.2 Contingency Table Analysis

In statistics, a contingency table (also referred to as cross tabulation or cross tab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. The degree of association between the two variables can be assessed by a number of coefficients: the simplest is the phi coefficient defined by where $\chi 2$ is derived from Pearson's Chi-Squared test, and *N* is the grand total of observations. It varies from 0 (corresponding to no association between the variables) to 1 or -1 (complete association or complete inverse association). This coefficient can only be calculated for frequency data represented in 2 x 2 tables. The phi (φ) can reach a minimum value -1.00 and a maximum value of 1.00 only when every marginal proportion is equal to 0.50 (and two diagonal cells are empty). Otherwise, the φ coefficient cannot reach those minimal and maximal values (Ferguson, 1966). The genotypes of a single, bi-allelic SNP on a set of cases and controls can be summarized in a 2×3 contingency table of the genotype counts for each group. For a SNP with alleles G and T, we tabulate the number of cases and controls with each genotype GG, GT, and TT. Several different statistical analysis methods can be applied to this table (Lewis *et al.*, 2009).

3.4.3 Positive and Negative Predictive Values

The positive and negative predictive values of a test depend, not only upon the sensitivity and specificity of the test, but also upon the prevalence of the disease in the population being studied. Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function.

Sensitivity measures the proportion of actual positives which are correctly identified as actual positive. In other words, it is the percentage of sick people who are correctly identified as having the illness. Specificity determines the proportion of negatives which are correctly identified. In other words, it is the percentage of healthy people who are correctly identified as not having the condition. These two measures are closely related to the concepts of type I and type II errors. A perfect predictor would be described as 100% sensitivity (i.e. predict all people from the sick group as sick) and 100% specificity (i.e. not predict anyone from the healthy group as sick), however theoretically any predictor will possess a minimum error bound known as the Bayes error rate.

Sensitivity judges a test's ability to identify positive results i.e. the proportion of people who have the disease and who have tested positive for it. This can also be explained as:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

Figure 3.2 Sensitivity Test Formula

Specificity judges a test's ability to identify negative results i.e. the proportion of people who do not have the disease and who will test negative for it. This can also be explained as:

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

Figure 3.3 Specificity Test Formula

In clinical diagnostics tests, the sensitivity test is the ability of a test to correctly identify those with the disease (true positive rate), whereas the specificity test is the ability of the test to correctly identify those without the disease (true negative rate). For example, if 100 patients known to have a disease are tested, and 65 test positive, then the test has 65% sensitivity. If 100 patients without the disease are tested and 92 return a negative result, then the test has 92% specificity. Sensitivity and specificity are prevalence-independent test characteristics, as their values are intrinsic to the test and do not depend on the disease prevalence in the population of interest.

The positive predictive value (PPV) is the proportion of persons with positive test results who are correctly diagnosed – i.e., the likelihood that a positive test result is correct in the context of the population being tested. The PPV can be calculated using the formula defined in Figure 3.4.

$$\text{PPV} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Positives}} = \frac{\text{number of True Positives}}{\text{number of positive calls}}$$

Figure 3.4 Positive Predictive Values Formula

Note that the positive and negative predictive values can only be estimated using data from a cross-sectional study or other population-based study in which valid prevalence estimates may be obtained. In contrast, the sensitivity and specificity can be estimated from case-control studies. If the sensitivity and specificity is known then the PPV can be calculated using the formula in Figure 3.5.

$$\mathrm{PPV} = \frac{(\text{sensitivity})(\text{prevalence})}{(\text{sensitivity})(\text{prevalence}) + (1 - \text{specificity})(1 - \text{prevalence})}$$

Figure 3.5 Positive Predictive Values if Sensitivity and Specificity are Known

The negative predictive value is the proportion of persons with negative test results who are correctly diagnosed – i.e., the likelihood that a negative test result is correct in the context of the population being tested. It is a summary statistic used to describe the performance of a diagnostic testing procedure. A high negative predictive value for any giving diagnostic test indicates that when the test yields a negative result, it most likely correct in its assessment. In a medical diagnostic analysis, a high negative predictive value assures that the test only rarely misclassifies a diseased person as being healthy. However, it does not indicate that the test will not mistakenly classify a healthy person as diseased.

$$\mathrm{NPV} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Negatives}}$$

Figure 3.6 Negative Predictive Values Formula

The negative predictive value formula if the sensitivity and specificity is known can be calculated using the formula in Figure 3.7. Although sometimes used synonymously, a negative predictive value generally refers to what is established by control groups, while a negative post-test probability rather refers to a probability for an individual. Still, if the individual's pre-test probability of the target condition is the same as the prevalence in the control group used to establish the negative predictive value, then the two are numerically equal. It is important to also note that the negative predictive value decreases with high prevalence of disease whereas the PPV increases with a high prevalence of disease.

$$\mathrm{NPV} = \frac{(\text{specificity})(1 - \text{prevalence})}{(\text{specificity})(1 - \text{prevalence}) + (1 - \text{sensitivity})(\text{prevalence})}.$$

Figure 3.7 Negative Predictive Values formula if the Sensitivity and Specificity are known

3.4.4 Area under the ROC Curve

The accuracy of a diagnostic test is measured by the AUC or the area under the receiver operator curve (ROC). An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. Sensitivity and specificity are the basic measures of the accuracy of a diagnostic test. They describe the abilities of a test to enable one to correctly diagnose disease when disease is actually present and to correctly rule out disease when it is truly absent. The accuracy of a test is measured by comparing the results of the test to the true disease status of the patient.

The ROC plot has many advantages over single measurements of sensitivity and specificity. The scales of the curve i.e., sensitivity and false positive ratio, are the

basic measures of accuracy and are easily read from the plot; the values of the cut points are often labelled on the curve as well. Because sensitivity and specificity are independent of disease prevalence, so too is the ROC curve. The curve does not depend on the scale of the test results (i.e., we can alter the test results by adding or subtracting a constant or taking the logarithm or square root without any change to the ROC curve). The ROC curve enables a direct visual comparison of two or more tests on a common set of scales at all possible cut points.

Diagnostic tests with ROC curve areas greater than 0.5 have at least some ability to discriminate between patients with and those without disease. The closer the ROC curve area is to 1.0, the better the diagnostic test.

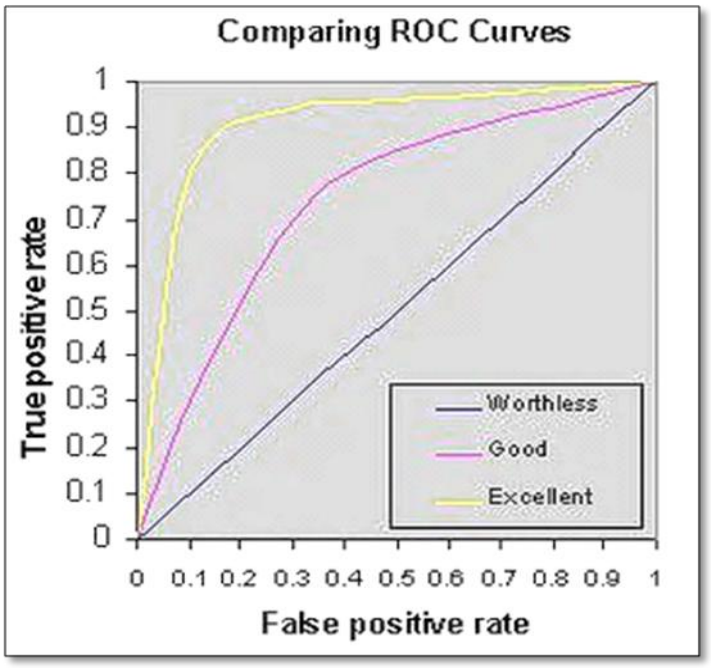

Figure 3.8 A Comparison of ROC Curves

3.4.5 Diagnostic Odds Ratio

The diagnostic OR is used to measure the effectiveness of a diagnostic test. It is defined as the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease.

Mathematically, the OR can be calculated using the formula shown in Figure 3.9. The rationale for the diagnostic OR is that it is a single indicator of test performance (like accuracy) but which is independent of prevalence (unlike accuracy) and is presented as an OR, which is familiar to medical practitioners.

The diagnostic OR ranges from zero to infinity, although diagnostic ORs less than one indicate that the test can be improved by simply inverting the outcome of the test. A diagnostic OR of exactly one means that the test is equally likely to predict a positive outcome whatever the true condition. Higher diagnostic ORs are indicative of better test performance.

$$\text{Diagnostic odds ratio} = \frac{TP/FN}{FP/TN}$$

Figure 3.9 Diagnostic OR Formula

3.4.6 Linear, Multiple and Logistic Regression

Statistical interaction can best be described in relation to a linear model that describes the relationship between an outcome variable and some predictor variable or variables. In linear regression, a model with a quantitative outcome ***y*** as a function of a predictor variable ***x*** using the regression equation

$$y = mx + c$$

is created. Here the regression coefficient ***m*** corresponds to the slope of the best-fit line and the regression coefficient ***c*** corresponds to the intercept. The use of the values of pairs of data points ***(x, y)*** (for example, if ***x*** and ***y*** are, respectively, measurements of height and weight in different individuals) to estimate ***m*** and ***c***, such that the line

$$y = mx + c$$

fits the observed data as closely as possible. In multiple regressions, this idea can be extended to include several different predictor variables using an equation such as

$$y = m_1x_1 + m_2x_2 + m_3x_3 + c$$

Here it is implicitly assumed that there is a linear relationship between each of the predictor variables, x_1, x_2 and x_3 and the outcome variable y, so that for each unit increase in x_1, y is expected to increase by m_1 (and similarly for x_2 and x_3). In logistic regression, rather than modelling a quantitative outcome y, the log odds,

$$Ln(\frac{p}{(1-p)})$$

in which p is the probability of having a disease, is modelled. For example, the model

$$Ln\left(\frac{p}{(1-p)}\right) = \alpha + \beta xB + \gamma xC + ixBxC$$

in which xB and xC are measured binary indicator variables that represent the presence or absence of genetic exposures at loci B and C respectively, β and γ are regression coefficients that represent the main effects of exposures at B and C, and the coefficient i represents an interaction term (a term that is required in addition to the linear terms for B and C). For two or more known or hypothetical genetic factors that influence disease risk, arguably the most natural way to test for statistical interaction on the log odds scale is to fit a logistic regression model that includes the main effects and relevant interaction terms and then to test whether the interaction terms equal zero. A similar approach can be used for quantitative phenotypes, in which case linear rather than logistic regression is used. These analyses can be

performed in almost any statistical analysis package after construction of the required genotype variables (Cordell, 2009).

3.5 Computational Tools Used

A number of computational programs/tools have been used to perform the analytical and data assessment tasks of this research project. For the benefit of readers, a brief overview of these software have been provided here. Due acknowledgement to the developers of these software must be made here for allowing the progress of our research through the use of these computational user-friendly applications.

3.5.1 The Plink Application Program

Plink is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. The focus of Plink is purely on analysis of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Since it has been written in C/C++, it can run in any operating system environment. It is readily available in DOS/Windows as well as UNIX/Linux environments. The Plink source code can be readily obtained and compiled for any particular operating system using any standard C/C++ compiler (Plink, 2009).

Through integration with gPlink and Haploview, there is some support for the subsequent visualization, annotation and storage of results. Plink is being developed by Shaun Purcell at the Centre for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others (Plink, 2009).

This toolset is able to perform data management tasks, calculate summary statistics for quality control, perform population stratification detection techniques, calculate basic association testing, identify multi-marker predictors and haplotypic tests, perform copy number variant analysis, gene based tests of association, screen for epistasis, identify gene-environment interaction with continuous and dichotomous environments, perform meta-analysis and also provide result annotations and reporting methods. Additional features of Plink include its capability to incorporate R function plug-ins, have web-based SNP and gene annotation lookup features, it has simple SNP simulation features, ID helper tools, to name just a few.

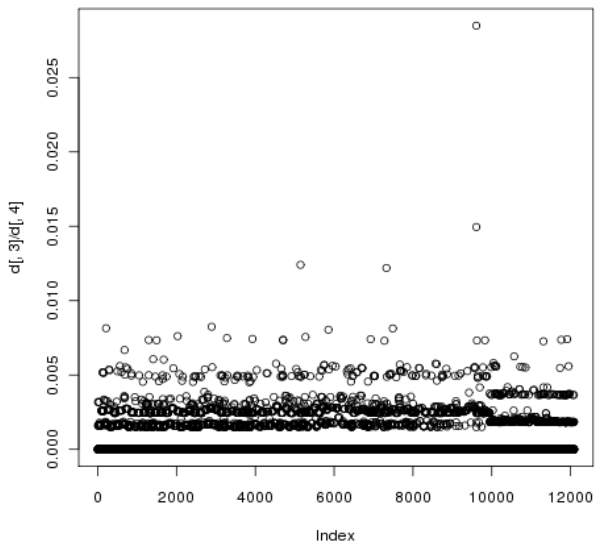

Figure 3.10 Plink Output Example

The vast features of Plink have allowed the use of this software for our specific data analysis purposes with great efficiency in its performance. Due acknowledgement must be made to the Plink developers for the extensive information that can be obtained from Plink resources, which allow users to properly grasp the best way to use Plink and provide a thorough understanding of the functions that are performed while using this tool.

3.5.2 The 'R' Statistical Application Program

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac-OS (R Project, 2012). R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland New Zealand, and now, R is developed by the R Development Core Team, of which Chambers is a member. R is named partly after the first names of the first two R authors (Robert Gentleman and Ross Ihaka), and partly as a play on the name of S (R Website, 2012).

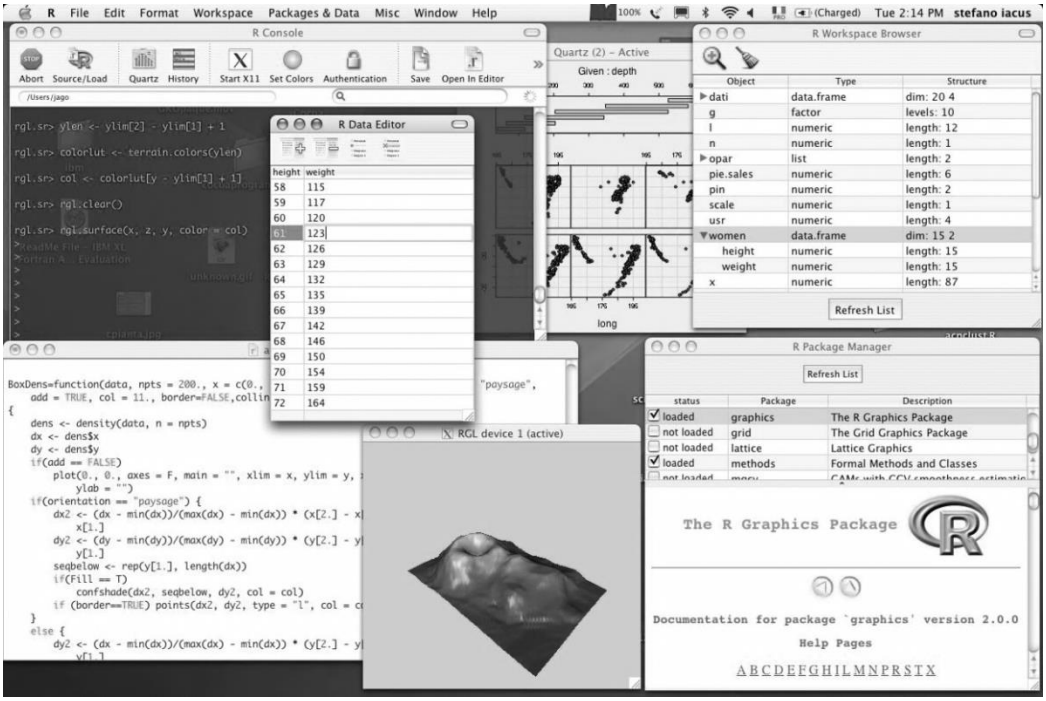

Figure 3.11 R Output Examples

The program “R” provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc.) and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac-OS.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either directly at the computer or on hardcopy, and
- a well-developed, simple and effective programming language (called ‘S’) which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R is very much a vehicle for

newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R program are essentially ephemeral, written for a single piece of data analysis.

The R environment is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities. The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

The software R is designed as a computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easier for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and FORTRAN code may be linked and called at run time. Advanced users can write C code to manipulate R objects directly. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics. R has its own LaTeX-like documentation format, which is used to supply comprehensive

documentation, both on-line in a number of formats and in hardcopy (R Project, 2012).

The R program was used for data analyses in parts of the Genomic Signature Analysis section of this research project. Due acknowledgement must be made to the developers of R for the extensive information that is provided that makes the learning and implementation of this program as user friendly as possible.

3.5.3 SPSS – Statistical Software

SPSS (originally, Statistical Package for the Social Sciences, later modified to read Statistical Product and Service Solutions) was released in its first version in 1968 after being developed by Norman H. Nie, Dale H. Bent and C. Hadlai Hull. SPSS is among the most widely used programs for statistical analysis in social science. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. The original SPSS manual has been described as one of "sociology's most influential books".

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the data file) are features of the base software. SPSS was released in its second version in 1972 and its company name is INDUS Nomi. The many features of SPSS are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses. Additionally, some complex applications can only be programmed in syntax and are not accessible through the menu structure. Programs can be run interactively or unattended, using the supplied Production Job Facility.

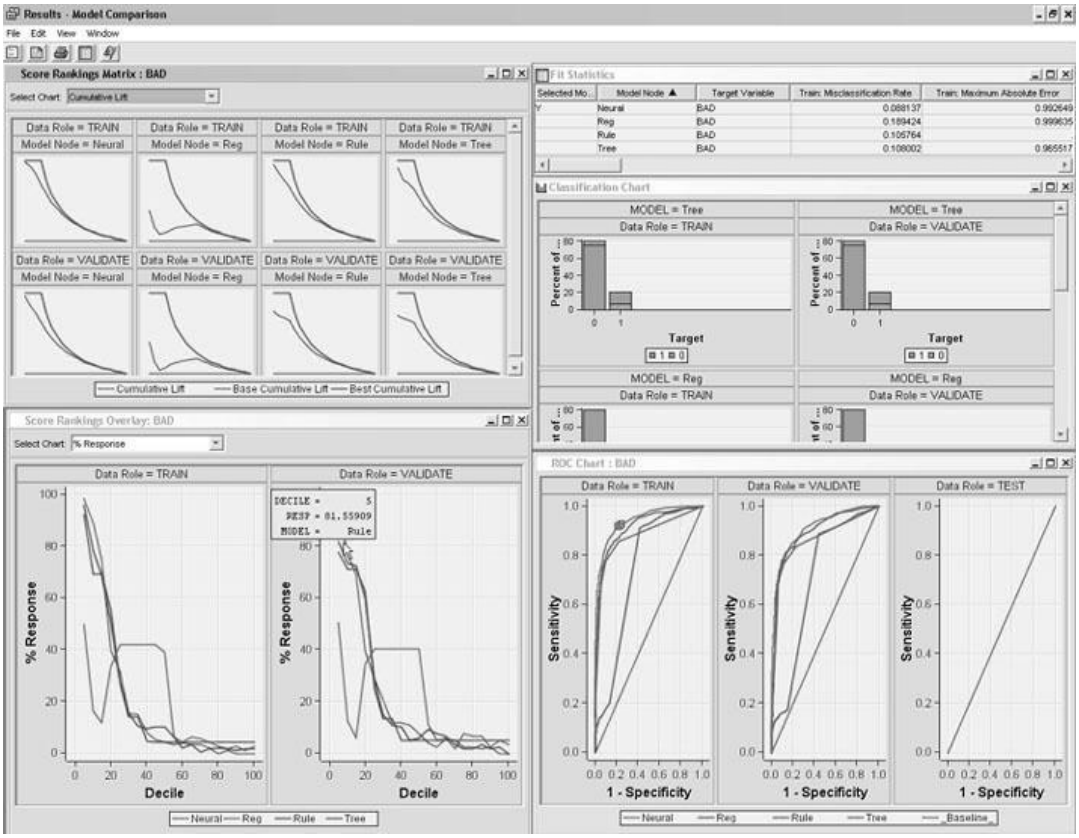

Figure 3.12 SPSS Output Examples

SPSS places constraints on internal file structure, data types, data processing and matching files, which together considerably simplify programming. SPSS datasets have a 2-dimensional table structure where the rows typically represent cases (such as individuals or households) and the columns represent measurements (such as age, gender or household income). Only 2 data types are defined: numeric and text (or "string"). All data processing occurs sequentially case-by-case throughout the file. SPSS can read and write data from ASCII text files (including hierarchical files), other statistics packages, spread sheets and databases. SPSS can also read and write to external relational database tables via ODBC and SQL codes (IBM, 2013).

The SPSS software was used extensively throughout each section of this research project. Sincere acknowledgements must be made for the detailed material that is

provided which as a result made the performance of comprehensive analytical methods which were applied as part of this research project more explicable.

3.5.4 The Multifactor Dimensionality Reduction (MDR) Software

The multifactor dimensionality reduction (MDR) software is open-source software available from http://www.multifactordimensionalityreduction.org/. The MDR is a data mining strategy for detecting and characterizing nonlinear interactions among discrete attributes (e.g. SNPs, smoking, gender, etc.) that are predictive of a discrete outcome (e.g. case-control status).

MDR is a non-parametric alternative to logistic regression for detecting and characterizing nonlinear genetic interactions. It facilitates data mining for detecting and characterizing combinations of attributes or independent variables that interact to influence a dependent or class variable. MDR was designed specifically to identify interactions among discrete variables that influence a binary outcome and is considered a nonparametric alternative to traditional statistical methods such as logistic regression. The MDR software combines attribute selection, attribute construction and classification with cross-validation to provide a powerful approach to modelling interactions. MDR is a nonparametric and genetic model-free data mining alternative to logistic regression for detecting and characterizing nonlinear interactions among discrete genetic and environmental attributes. The MDR method combines attribute selection; attribute construction, and classification with cross-validation and permutation testing to provide a comprehensive and powerful approach to detecting nonlinear interactions (Moore *et al.*, 2013).

The basis of the MDR method is a constructive induction algorithm that converts two or more variables or attributes to a single attribute. This process of constructing a

new attribute would change the representation space of the data. The end goal is to create or discover a representation that facilitates the detection of nonlinear or non-additive interactions among the attributes such that prediction of the class variable is improved over that of the original representation of the data. The MDR software collapses high-dimensional genetic data into a single dimension thus permitting interactions to be detected in relatively small sample sizes. MDR is a promising new approach for overcoming some of the limitations of logistic regression for the detection and characterization of gene–gene and gene–environment interactions (Hahn *et al.*, 2003).

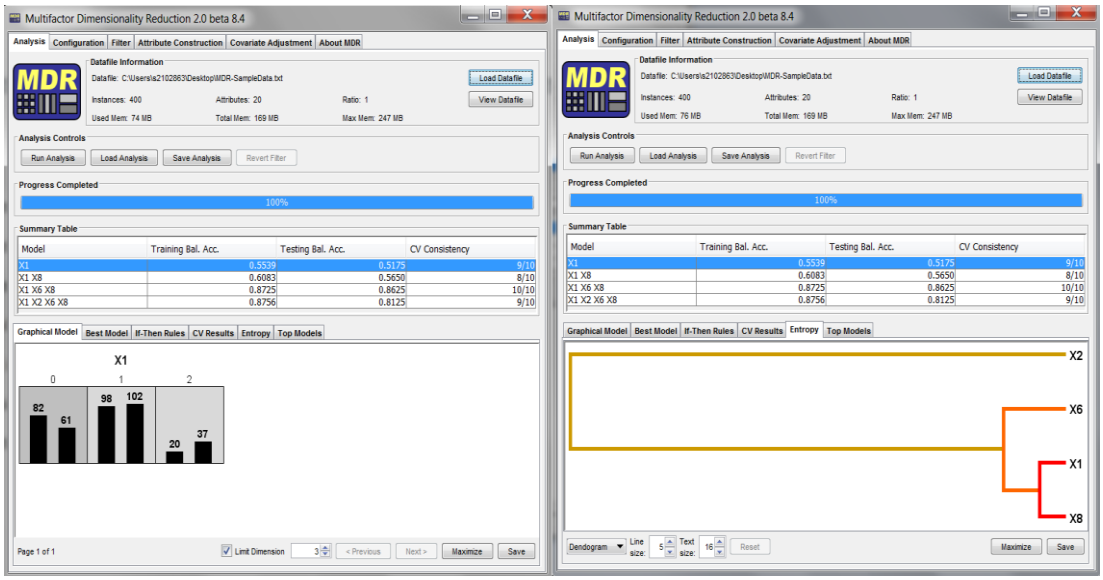

Figure 3.13 MDR Output Examples

MDR analysis was performed as part of the Genomic Signature Analysis research methods of this project. A genuine acknowledgement must be made to the developers of this comprehensive tool for the extensive information and user help that is provided along with this free and readily available application. Performing our research work has been made possible only because such tools are freely available for one and all, and extensive information is provided to completely and successfully use this tool.

3.6 Summary

CD is a chronic disease with a substantial impact on a patient's quality of life. The care of CD is challenging as the diagnosis is often delayed, and there are a multitude of treatment options to consider. Many of the medications for CD are immunosuppressive, thus vigilant follow-up and laboratory monitoring is required to prevent and/or minimize complications. The severity and progression of disease is variable, most often manifesting a relapsing/remitting pattern over the patient's life span, and may require both medical and surgical intervention. There are important health maintenance considerations for these patients, including the provision of vaccinations and screening for osteoporosis, colon cancer, and vitamin deficiencies. Patient education and close coordination between primary care providers and consultants is critical in achieving positive outcomes in these often complicated patients (Walsh *et al.*, 2011).

Several studies have shown that the implementation of biologic therapies in active refractory CD is associated with mucosal healing and a reduced number of hospitalizations and surgeries. Previous studies have shown that clinical and demographic markers alone poorly predict disease evolution. This suggests that an early introduction of such therapies might prevent a disabling disease course. Therefore it becomes increasingly desirable to define patients at high risk for such an evolution in an early stage (Henckaerts *et al.*, 2009).

In February 2011, more than 10 years after a draft sequence of the human genome was published, the US National Human Genome Research Institute announced its new strategic plan for genomic medicine from base pairs to bedside. The plan calls for evaluating the structure and biology of genomes; understanding the biology of disease; advancing the science of medicine; and improving the effectiveness of

healthcare. Nevertheless, fulfilling the promise of genomics in improving health requires a multidisciplinary research agenda beyond bench to bedside, an agenda that will demonstrate added value of genome-based information for improving health in populations.

Currently, this translational research agenda is much less robust than discovery research, accounting for less than 2% of funded genomics research and research publications, but this is likely to change in the next decade as more and more applications make it to the bedside. Ongoing horizon scanning for genomic tests has revealed more than 330 genomic tests have reached the bedside since 2009. With rapid improvements in technologies, we are seeing the leading edge of the applications of WGS in practice primarily in the identification of rare heritable conditions. The ultimate success of genomics for improving health will require adoption of evidence-based approaches for their use in clinical and public health practice (Auffray *et al*, 2012).

Consequently, in order to completely utilise all scientific knowledge it is important to tackle large-scale research projects with open collaborations so that a multi-disciplinary approach can be undertaken. This would subsequently allow the merging of many faculties of science, and nurture the use of many scientific minds, each with their own detailed ideas and subject knowledge. With the utilisation of numerous interacting fields of science, the benefits that can be achieved for the consumer and/or patient at the end would be remarkably noteworthy.

CHAPTER FOUR: ENVIROGENOMIC RISK PROFILING

Acknowledgement: This chapter includes co-authored papers that have been either already published or are under review for publication, with myself as the main author. Further details regarding the papers are provided in the section titled "List of Publications and Conference Presentations".

4.1 Introduction

It is becoming increasingly evident that single-locus effects cannot explain complex multi-factorial human diseases like CD (Manuguerra *et al*, 2007). Treatment of these complex diseases has not seen much improvement based on any one factor influencing the disease. Development of CD is believed to be as a consequence of multi-factorial interactions between genetic, immune-related, and environmental factors (Economou *et al*, 2008). It makes sense therefore to deduce that treatment techniques can benefit immensely if all possible disease risk factors are analysed in a collective mode.

When the single polymorphism effect is not present alone or is not strong enough, the identification and characterization of susceptibility genes for risk require the understanding of gene-gene (Manuguerra *et al*, 2007) and gene-environment interactions. Understanding these gene-environment interactions will be able to provide a more precise synopsis of the characteristics of a disease and ultimately allow for a more accurate diagnostic analysis as well. Ultimately, if patients can be diagnosed more accurately for their risk of disease at the very start of treatment i.e. at the time of diagnosis, it can be assumed that they will be provided with a more

specific treatment regimen that will be able to treat the disease and prevent the occurrence of complicated progressive forms of disease.

The majority of scientific research is not geared towards prediction of clinical outcomes but rather focuses on better understanding disease pathology. Only recently, it has been seen that a small amount of emphasis has begun to emerge on the importance of incorporating environmental and other non-genetic factors into disease risk prediction analysis. Methods investigated by Chatterjee & Wacholder (2008) and those explained by Thomas (2010), as well as others, all realise that GWAS alone cannot predict real-time risk of complex diseases.

Understanding the role of interactions between genetic and non-genetic factors on the relative risk of causing disease in individuals has many advantages. As a result of analysing gene-environment interactions to create risk prediction models, both the overall health of the community and that of an individual, can achieve beneficial outcomes. This research on envirogenomic risk profiling aims to identify envirogenomic signatures that accurately predict a complicated outcome in CD patients by analysing genetic, clinical and environmental data in combination. This means that all genetic, clinical and environmental factors, each with their own individual influence towards disease risk, will be analysed en masse so that an overall combined disease risk value can be determined for each patient. This practical approach to analysing factors of disease creates a more personalized predictive model that can be used for diagnostic purposes and be tailored for each individual by clinicians.

Reliably foretelling the progression and clinical outcome of CD is quite complex. The disease displays many characteristics varying between each individual and it is

quite difficult to determine a causative factor for a specific individual at the time of diagnosis. Several clinical predictors have been previously identified including active smoking, young age at onset, age at diagnosis, extensive small and/or large bowel disease, need for steroids at diagnosis, extra intestinal manifestations and perianal disease (Henckaerts *et al*, 2009).

Genetic markers have also been explored as predictors of disease outcome and suggested as advantageous because the genetics of a patient can be analysed before any sort of clinical or environmental factors start playing a role in disease progression (Henckaerts *et al*, 2009). Many genetic studies of clinic-based patient cohorts have been performed to date and variants in the NOD2 gene have been frequently associated with complicated characteristics including shorter-time to stenosing disease, development of structures and fistulae, as well as need for surgery (Vermeire *et al*, 2007).

Some CD patients will develop complicated disease behaviour characterized by penetrating and stricturing symptoms which can cause a decrease in the quality of life and in some cases the need for surgery (Henckaerts *et al*, 2009). Overall, patients suffering from IBDs such as CD will show symptoms of disease that affect their daily well-being and these symptoms range from a vast array of causes including genetic, environmental and clinical risk factors. These risk factors all indicate causation for disease progression to complicated forms if left undiagnosed and untreated. Hence, it is essential for clinicians to identify each and every risk factor possible at the time of diagnosis to allow for a precise and clear indication of treatment choices.

CD usually adopts a natural course which involves periods of remission in between flare ups of disease symptoms. Treatment usually depends on controlling disease symptoms and maintaining remission, however a majority of CD patients end up having surgery within the first ten years of diagnosis (Vermeire *et al*, 2006). Current treatment therapies rely on two known approaches: the top down and the step up approach, and have evolved from the impact of novel anti-TNF therapies (Hanauer *et al*, 2003). These two treatment approaches differ depending on disease severity of patients and evidence (discussed previously) has shown that early aggressive treatment is beneficial for complicated disease cases in preventing the need for surgery, instead of the step wise sequential treatment approach which is more beneficial for less severe disease symptoms (Hanauer *et al*, 2003). Therefore the subgroup of CD patients at risk of developing "complicated" CD might benefit from intensive early treatment with biologicals (e.g. TNF-alphas) – the so-called top-down approach (Baert *et al*, 2007).

A European study provided evidence that genetic risk profiles might be useful for predicting complicated outcomes in a clinic-based cohort of CD (Henckaerts *et al*, 2009). However, there are many environmental and clinical factors that lead to the development and progression of CD, and it is necessary to assess these factors along with an individual's genetic risk as well, to determine a more comprehensive and precise assessment of the actual risk of complicated disease development.

Genetic risk factors are predetermined and cannot be controlled or altered by a patient, however a patient is able to alter their environmental influences, and knowing which environmental factors are more of a risk for a patient might benefit the patient and prevent the risk of developing complicated disease forms. Even though gene-environment interaction studies are much more difficult to analyse than

pure genetic associations, the important role of environmental factors should not be ignored when attempting to predict disease risk. For complex multi-factorial diseases such as CD, where the genetic impact on disease is relatively low, the knowledge of other disease causing risk factors is immensely valuable in increasing complicated disease risk prediction measurements and thus creating beneficial prognostic tests.

Because of the many differences in the genetic, environmental and clinical characteristics of CD, it means that patients tend to respond differently to the types of treatments currently available and prescribing optimal types of treatment can be a real challenge for gastroenterologists. Usually treatments are prescribed based on diagnosed manifesting symptoms but fail to acknowledge underlying causes that may be present but have not yet been observed. It is an enormous burden for clinicians or gastroenterologists to assume how a disease will progress based on these unknown or invisible factors and assess the future risk of complicated disease development for any given individual patient. Thus, this research aims to lighten the burden of clinicians by eventually providing an accurate diagnosis prediction model that can be used in to create a practical clinical tool to aid in decision making.

The importance for clinicians to be accurate in their diagnosis and risk prediction of patients when prescribing individual treatments is emphasized due to the many side effects and the financial burden of the top down approach of treatment. Usually medical practitioners refrain from prescribing the top-down treatment approach for treating CD at the start of diagnosis, because they have no knowledge of a patient's risk of progression of complicated disease. This highlights the need for better and more accurate prognostic and diagnostic tests for predicting clinical outcomes for CD patients. In this study, the interactions of genetic, clinical and environmental factors, both independently and in combination, to clinical outcomes that are prevalent in a

vast number of CD patients is investigated. It is highly probable that the predetermined hereditary genetic variants when considered in combination with clinical and environmental factors would be able to present improved clinical diagnostic outcomes, more accurate and specific than when any factor is considered independently.

This research project has obtained data sets from a number of different organizations/groups and has performed a couple of varying research methods on this data. The first section of the first part of the research study looks at analysing genetic and clinical data from a population-based CD cohort from New Zealand in an effort to identify personalized risk profiles associated with increased risk of complicated clinical outcomes and which might be useful as prognostic tests. This part of the study aims to clarify the risk of the need for surgery as a clinical outcome for CD patients based on several environmental, clinical and genetic factors when analysed in combination. It aims to obtain an accurate and clinically useful value for the prediction of the risk of surgery in CD patients. For the second section of this part of the study it is hypothesized that a *combined* factor analysis using genomic, clinical and environmental factors will provide predictive personalized profiles for the early prediction for the need of surgical intervention in CD patients.

4.2 Specific Objectives

The objective of the envirogenomic risk profiling section of this research project was to perform a case-control study of CD patients from the Canterbury IBD cohort for identifying envirogenomic risk profiles that predict the need for surgery as a clinical outcome of CD. Specifically, the study was divided into two parts, each with separate distinct objectives. The first objective was to perform a case-control study of

~700 CD patients from the Canterbury IBD cohort to identify envirogenomic profiles of complicated CD. This analysis for this objective was performed in both a retrospective approach as well as in a prospective manner. The second objective was to replicate identified envirogenomic profile associations in an independent CD cohort collected from Australia and New Zealand.

The analytical techniques applied to the data for the two separate research investigations involved separate examination approaches which have been explained herewith. The analytical techniques are described in a step wise fashion exactly as they were performed on the data during the process of this research investigation to allow a better understanding of the undertaken research methodologies.

4.3 The Canterbury IBD Study

In 2006 a paper based on the Canterbury IBD Project was published that enforced the fact that an increasing trend in the number of individuals with CD was being witnessed in Canterbury, New Zealand (Gearry *et al*, 2006). Specialists working in this region noticed that there were an increasing number of patients being diagnosed for CD in the Canterbury region specifically over recent years. As a result, a comprehensive study was undertaken by Dr. Richard Gearry over the course of three years of his PhD to gain more knowledge and data of the CD trends in the Canterbury region.

The Canterbury IBD Study was a population-based research project designed to investigate IBD epidemiology (frequency), aetiology (causes), clinical outcomes and treatments. Over 1400 IBD patients (comprising more than 92% of all people with IBD in Canterbury in 2005) were recruited into the study. Each subject completed a questionnaire and gave blood samples for genetic and other clinical analyses.

Collaborators on the IBD project included those from The Gene Structure and Function Laboratory of the University of Otago, Christchurch, the Nutrigenomic group in Auckland and the Crohn's and Colitis Support Group of New Zealand.

From this study, it was established that IBD was at least as common in Canterbury as in other Western regions and that the incidence and prevalence of CD were amongst the highest ever reported in the world (Gearry *et al*, 2006). Patient population characteristics were otherwise similar to other demographically matching countries (Gearry *et al*, 2006). The study revealed that in 2004, age-standardized (World Health Organization World Standard Population) IBD and CD incidence rates were 25.2 and 16.5/100,000/year, respectively in Canterbury. The IBD and CD point prevalence's on 1 June, 2005 were 308.3 and 155.2/100,000, respectively. CD patients were more likely than UC patients to be female (61.4% vs. 47.1%) and to be younger (median age, 39.9 years vs. 43.7 years). The percent of IBD patients who were white was 97.5% (Gearry *et al*, 2006). A vast amount of data ranging over a significant number of factors was collected as part of this study. The data collected included a range of environmental, clinical as well as genetic factor information. (More important statistical information about this study by Dr Gearry can be obtained from his publications using this population). As a result, the data collected as part of the Canterbury IBD Project was extremely relevant and held importance for the research that was required to be carried out for this project.

4.3.1 Participants

In order to recruit patients for this study, a number of factors were considered and kept in mind to create a patient cohort that would provide accurate and universally uniform data which could be classified as significant and relevant for this study.

Unlike countries with universal health systems that have meticulous coding and diagnostic or therapeutic registry data, New Zealand has no such system (Gearry *et al*, 2006). Although most New Zealand residents use the public health system, 14% have comprehensive private health insurance, and therefore patients attending both public and private clinics and hospitals need to be identified for a comprehensive study (Gearry *et al*, 2006).

All people with IBD living in the Canterbury region, irrespective of their age, gender or disease severity were included as part of the study. A total of 1420 Cantabarians with IBD gave informed consent to take part in Gearrys' initial study. Each of these individuals provided a blood sample for DNA, completed a questionnaire regarding environmental factors and gave permission for researchers to evaluate their medical history records. This information was stored in secure computer databases. Individuals without IBD that were used as controls also provided the same information to allow for comparisons and appropriate statistical evaluations.

Specifically, for the purposes of this research project, a total of 709 patients with CD from the Canterbury IBD study were extracted for analytical research purposes, and from these 709 patients, there were 503 that had provided genotype data available for analysis. Only patients with complete genotype data available for analysis were used for the purposes of this research project. From these 503 patients, dichotomisation into cases and controls was carried out based on the primary outcome variable that was analysed in this study. Statistical power calculations from this dataset estimated a 93% power of test based on the sample size utilized.

4.3.2 Data Collection

Cases for this study were recruited in multiple ways, including patient advertising, letters to patients from their doctors, and approaches to patient support groups. To help with the recruitment of incident cases prospectively, gastroenterologists, surgeons, and general practitioners were informed of the study both by letter and meetings and kept up to date. The investigators were then informed about the patient, and their diagnosis and patient consent was obtained before data was collected from the IBD patients (Gearry *et al*, 2006).

Cases comprised patients with an established diagnosis of CD, confirmed by case note reviews using established criteria by a gastroenterologist with an interest in IBD (Gearry *et al*, 2010). Cases not meeting diagnostic criteria, without acknowledged consent, participation decline, or not living in the Canterbury region were excluded. All patients were also phenotyped according to the Montreal and Vienna classification systems (Gearry *et al*, 2006). The study gained ethical approval from the Canterbury Ethics committee and written informed consent was obtained from all subjects prior to data collection (Gearry *et al*, 2006).

Patients were allocated a four digit study number to ensure anonymity, and were given an environmental questionnaire and underwent vene puncture to collect genetic data (Gearry, 2006). A self-administered questionnaire was developed to determine the presence, absence and timing of exposure to many environmental factors and was trialled in 20 IBD patients and 20 controls before using it for the entire cohort. A total of 102 environmental association factors were examined (Gearry *et al*, 2010). Some of the important environmental data that was recorded from the self-administered questionnaire is outlined in Figure 4.1.

Demographics	**Miscellaneous**	**Childhood factors**
Date of birth	Current income	Place of childhood
Sex	Current occupation	Exposure to farm animals
Country/place of birth	Vegetarianism	Exposure to Pets
Ethnic group	Current alcohol intake	Person : bedroom ratio
Descent from Maori	Non-prescribed treatment	Number of toilets
Educational achievement	Onset of symptoms (date)	Bedroom sharing
Family	Diagnosis date	Vegetable garden
Sibship size and rank	Public swimming pool use	Takeaway food frequency
Maternal age at birth	Attendance at preschool	Antibiotics
Family history of IBD	**Smoke exposure**	% of house carpeted
Family history (other illness)	Personal smoking history	Presence of a sandpit and
Neonatal Factors	Childhood smoke exposure	swimming pool
Breastfeeding	Maternal smoking	Method of home heating
Duration of breastfeeding	**IBD morbidity**	Antibiotic use
Maternal medications	Disability	Regular medications
Paternal occupation	Absenteeism	**Medical history**
Immunisation	Fecundity (female)	Medical conditions
		Previous surgery
		OCP use

Figure 4.1 Environmental Data Recorded From the Self-Administered Questionnaire (Gearry, 2006).

Case recruitment was greater than 90% so population based controls from a common study base were used (Gearry *et al*, 2010). Controls were randomly selected from the Electoral Roll for electorates corresponding with the Canterbury District Health Board geographical boundaries (97% of New Zealanders are registered on the Electoral Roll) (Gearry *et al*, 2010).

4.3.3 Data Analysis

Clinical and environmental data was gathered for 709 CD patients in the Canterbury IBD study and genotype data was available for 503 patients. Eight variants from seven genes and three variants of the NOD2 gene were genotyped for the 503 patient case studies with CD from the Canterbury IBD cohort. All three variants of the NOD2 gene were summarized as one data column, therefore patients with any of the three variants were considered as having the NOD2 gene. A total of 113 factors were

extracted and analysed as part of this study. The 113 factors consisted of 8 genetic, 2 clinical and 102 environmental factors.

The primary outcome variable for this study was the need for IBD-related surgery in CD patients. Those patients who had previously undergone IBD-related surgery were classified as 'cases' and those who had not had any type of IBD-related surgery were classified as 'controls' for this study. A total of 241 cases had surgery within this cohort and 468 had never had IBD-related surgery. From the total case control group of 709 patients, a total of 70% *(n = 503)* patients were available for genotyping. Only those patients with genotype, environmental and clinical data available were included for the purposes of this study. Those patients with missing data were excluded.

To analyse the data that was available, two different investigation methods were applied. Firstly, a retrospective approach using common known predictive risk factors for the need of IBD-related surgery that have previously been identified in literature were extracted from the whole dataset and analysed. Secondly, the whole dataset in its entirety was prospectively examined and a step-wise data reduction approach was applied to determine risk factors that could be associated with IBD-related surgery in this specific population cohort.

4.3.3.1 Retrospective Analysis Methods

Study design, patient diagnosis and predictor variables:

A retrospective case control data analysis was performed using variables previously identified as causative for any type of IBD-related surgery in past research studies after a thorough literature review of CD risk factors. Those factors that had complete data available and were well known in literature as

being proven CD risk factors were extracted from the whole data set for this retrospective analysis. In brief, data was available for 709 patients diagnosed for CD using the Montreal Classification System.

The average age of the patients was 45 years, and the average disease duration was 9 years (no subsequent follow-up data was available at the time of this analysis). Selected clinical predictor variables included, age at diagnosis (grouped according to the Montreal Classification System as: 16 years or younger, 17-40 years, older than 40 years), active smoking at diagnosis and any type of perianal disease behaviour (i.e. patients with one or more manifestations of perianal abscesses, significant anal skin tags, fistula-in-ano, rectovaginal fistula, anal fissures or anal canal stenosis). These variables were selected because **(*a*)** they had been previously shown to predict CD in this cohort (Gearry *et al*, 2007 & Tarrant *et al*, 2008), **(*b*)** they are generally considered to be key risk factors for CD development (Vermeire *et al*, 2007), and **(*c*)** because data for these variables was complete and readily available for this analysis.

In addition, genotype data for three variants (R702W, G908, 1007fs) of the NOD2 gene were available for a subset of 503 patients collected as previously described (Gearry *et al*, 2006). Patients were classified as having the NOD2 gene if they had *any* of the three variants genotyped, and only those patients with *any* NOD2 information were included for this analysis. Treating clinicians were not aware of patient genetic information at the time of recruitment, and therefore NOD2 patient data is unbiased. Any patients with missing or incomplete data were removed from this data subset. Any other genetic information regarding patients was not available for analysis.

Clinical Outcome:

The need for CD-related surgery was considered as the clinical outcome for this study. This outcome was broadly defined as any type of intestinal bowel resection which occurred as a direct result of CD at the time of recruitment and classification was confirmed by a clinical gastroenterologist (Dr.Richard Gearry) via patient records. Patients were assigned dichotomous labels for the outcome i.e.

0: patients had no IBD-related surgery at the time of recruitment, or

1: patients had at least one IBD-related surgery at the time of recruitment.

Those patients who had no IBD-related surgery at the time of recruitment were classified as controls and those patients who had at least one IBD-related surgery at the time of recruitment were classified as cases for this study.

Statistical Analyses:

Multi-factorial logistic regression was performed using the need for surgery as the primary outcome variable. Several predictor variables, including NOD2, current age, gender, active smoking, perianal disease and age at diagnosis, were analysed. Stepwise forward conditioning was performed to create regression models that showed significant risk factors for complicated CD outcomes.

Using the significantly associated genetic information and clinical data, the NOD2 gene variable was then combined with the perianal disease variable to create a novel combined genetic and clinical variable (NOD2+PA) for analysis. This new NOD2+PA clinicogenetic variable was also modelled by logistic regression methods. Diagnostic statistics, including the calculations of the PPV

and the Attributable risk (AR) for this variable were also calculated for assessing its prognostic value.

4.3.3.2 Retrospective Analysis Results

The clinical characteristics of the CD patient cohort are shown in Table 4.1. Patients requiring IBD - related surgery were considered as patients with complicated disease outcomes for this study and termed as 'cases'. Those patients not requiring IBD-related surgery were considered as without complicated disease outcomes for this study, and termed as 'controls'. Regression models were created in which the NOD2 genotype, age at diagnosis, perianal disease, and active smoker at diagnosis predictors were analysed as predictor variables to assess significance within the cohort as a whole.

To assess our original objectives for this study, *i.e*. to analyse predictor variables in combination to evaluate their associations with complicated outcomes of CD, a combined variable model was also created. The combined regression model showed that NOD2 and perianal disease were significant predictors of the need for surgery in this patient group yielding odd ratios of 1.60 and 2.84, respectively (Table 4.2). All results were adjusted for age and sex accordingly.

Considering these substantial results, we combined the presence of the NOD2 genotype and perianal disease into a single predictive clinicogenetic factor (NOD2+PA). Analysis against need for surgery showed that the NOD2+PA clinicogenetic factor yielded an OR of 3.84 (*95% CI: 2.28-6.46*).

Table 4.1 Patient Characteristics for the Canterbury IBD Study

Characteristic	**Total**
Total Patients[I]	709 ($p < 0.001$)
Mean Age (SD) (%)	43 (18)
Need for surgery (%)	240 (33.9)
Active Smoker at diagnosis [II] (%)	361 (50.9)
Perianal disease (%)	189 (26.6)
Age at Diagnosis	
16 or younger (A1) (%)	76 (10.7)
17-40 (A2) (%)	386 (54.4)
40 and above (A3) (%)	206 (29)

[I] Patients include both cases and controls

[II] Smoking information only available for n=667/709 patients

Diagnostic statistics were then calculated using the new combined clinicogenetic NOD2+PA variable. Results showed a low sensitivity of 0.25 but a high specificity of 0.92 translating into a moderate PPV of 62%. The population attributable risk was 32%, suggesting that if the NOD2+PA variable were to be removed, while all other factors remained unchanged, the number of complicated surgery cases would decrease by 32%.

Table 4.2 Results of the Multi-Factor Logistic Regression Analysis for the Retrospective Canterbury IBD Study

Variable	B	SE	P-VALUE	OR	95% C.I.	
					Lower	Upper
Current Age	0.01	0.01	0.24	1.01	1.00	1.02
Gender	-0.11	0.21	0.49	1.16	0.77	1.74
Active Smoker	-0.13	0.21	0.53	1.14	0.76	1.70
PA disease	1.04	0.22	**0.00**	2.84	1.83	4.38
Age at diagnosis	0.01	0.01	0.40	1.01	0.99	1.02
NOD2 genotype	0.47	0.20	**0.02**	1.60	1.08	2.38

B: Regression coefficient, S.E: standard error, OR: Odds ratio, CI: confidence interval, PA disease: Perianal disease

4.3.3.3 Retrospective Analysis Discussion

This research study aimed to investigate the impact of a clinically relevant outcome of CD when predictive disease risk factors are evaluated not only individually, but also in combination. Studies have shown that early treatment for CD can reduce the number of surgeries and hospitalizations and quicken patient inflammatory healing (Vermeire *et al*, 2007). These studies have revealed that an early implementation of proper medical management of disease can also influence the path of disease and consequently most likely reduce the risk of complicated disease progression (Henckaerts *et al*, 2007). Therefore, it is important to identify patients that are more at risk of developing a complicated form of disease at the time of diagnosis to abstain from possibly avoidable medical interventions.

There are currently a range of non-specific medications available for the treatment of CD, but due to differences in individual symptoms that develop from the broad spectra of CD characteristics, patients tend to respond differently to individual methods of treatment. Deciding on the most effective drug treatment for each patient,

i.e. personalizing treatment can be a real challenge for gastroenterologists and highlights the need for better prognostic tests for predicting the clinical outcomes for CD patients.

This study investigated genetic and clinical data from a population-based cohort from the Canterbury IBD project. Several genetic and clinical studies have already been published on this cohort (Henckaerts *et al*, 2007, Roberts *et al*, 2007, Roberts *et al*, 2010, Hollis-Moffatt *et al*, 2010, Eglinton *et al*, 2010) but none so far have examined need for surgery as a clinical outcome or looked at combined factors and their interactions.

This study identified the NOD2 genotype as a significant risk factor for the need for surgery. These results are consistent with results from several other studies (Manuel *et al*, 2005, Laghi *et al*, 2005, Russell *et al*, 2005, Seiderer *et al*, 2006, Kugathasan *et al*, 2004) where the NOD2/CARD15 variants were associated with early initial surgery due to stenosis and with surgical recurrence in CD. In addition, we identified perianal disease as a predictor of surgery in this cohort, which has not been previously directly linked as being a predictor of the need for surgery for CD patients but has instead been identified as a predictor for behaviour change in CD instead (Lakatos *et al*, 2009). A recent study also has looked at a combination of genetic, clinical and immune markers to predict the need for surgery in CD patients (Dubinsky *et al*, 2011), but they assessed each of these predicting factors independently of each other and failed to assess a combined variable approach as is performed here.

Most importantly, this study showed that combining NOD2 and PA into a single risk factor achieves much higher associations than when either factor is considered

independently. Studies have shown that by combining genetic information associated with CD, the risk of the development and severity of CD can be predicted (Weersma *et al*, 2009), studies where clinical and genetic data have been combined together to assess the risk of complicated disease and surgical intervention have not been seen previously.

This research identified coupled associations between selected genetic and clinical factors that influence the risk of disease severity based on their clinically evaluated outcomes. To validate our study and prove its importance, firstly, our OR is quite high indicating that there is significant association between our predictors and the resulting outcome. Secondly, our relatively high specificity and low sensitivity indicate that the test is accurate and showing appropriate results. Also, based on our disease prevalence within the cohort, our PPV and NPV is also quite high indicating the accuracy of diagnosis of cases at risk for this variable and that they have been correctly identified.

Clinically, our study indicates those patients who have had IBD surgery, have the NOD2 gene, and also have perianal disease have more than 60% probability of developing complicated disease. Therefore, anyone with a similar diagnosis would be considered as having the same risk of complicated disease. While our results indicate significant diagnostic value, the PPV of 62% is clinically low to be used as a prognosis for surgery. This therefore indicates the necessity for further analysis with more genetic, clinical and environmental variables that can improve the prognostic value of this research.

Potential limitations of this analysis include the unavailability of disease behaviour and location data as well as the lack of information regarding immunosuppressive

treatment before surgery. Follow-up data for the variables analysed in this study were also unavailable at the time of this research analysis. These variables have been shown to impact the outcome of the need for surgical intervention for CD patients in past studies (Lakatos *et al*, 2012, Ramadas *et al*, 2010), therefore despite these potential limitations, the results obtained from this analysis are considerably significant and hope to provide beneficial insight into evaluating risk factors that predict clinical outcomes in CD patients.

From this study we can conclude that patients with disease predictors that cause a higher risk of surgery, when medicated with an aggressive top down treatment approach, may be able to reduce severe disease outcomes and decrease their risk of surgical intervention. A beneficial disease outcome such as long term remission or improved mucosal healing may lead to decreased disease progression, but this study has shown that a personalized aggressive therapeutic strategy for CD patients at diagnosis is the best way forward in the long term, in order to decrease the risk of surgical intervention and the problems they create for CD patients.

At present there are no efficient methods and systems that allow the application of personalized medicine to the delivery of health care in CD patients. It is hoped that these types of clinicogenetic tools will lead to prescribing patients with a more personalized action strategy, based on what treatments are more likely to be effective for the patient or may cause undesirable clinical outcomes.

4.3.3.4 Prospective Analysis Methods

Patient selection and Data Ascertainment:

The same case-control population that was used for the retrospective analysis methods was also used for this prospective analysis method. Hence, patient selection and data ascertainment is the same as the retrospective analysis method, however they will be repeated here for convenience and better understanding. Clinical data was gathered for 709 CD patients in the Canterbury IBD study. Genotype data was available for 503 patients. All 503 patients were used for all factor analyses. Eight SNPs for seven genes and three variants of the NOD2 gene were genotyped for the 503 patient case studies with CD from the Canterbury IBD cohort. Genotyped patients included 306 females and 197 males. The average age of the patients was 45 years, and the average disease duration was 9 years (no subsequent follow-up date was available at the time of analysis). Age and sex factors were adjusted to deal with any differences that may have risen during analysis.

Disease behaviour was classified using the Montreal Classification System. 160 patients were diagnosed with stricturing disease behaviour, 57 patients with penetrating and 286 patients with non-stricturing non-penetrating disease behaviour. A total of 211 patients had colonic disease location, 166 with ileal location, and 126 with ileocolonic disease location. A total of 63 variables including genetic, environmental and clinical factors were analysed as part of this study.

The primary outcome variable was the need for any type of IBD-related surgery in CD patients. A total of 174 (35%) cases had at least one IBD-related

surgery within this cohort and 329 (65%) had never had IBD-related surgery at the time of recruitment.

***Statistical Analysis**:*

A step-wise data reduction approach to analysing the data with the aim of identifying "envirogenomic" profiles that predict risk of surgery in CD patients was analysed. The term 'envirogenomic' aims to define a single variable that is created by combining environmental, clinical and genomic factors.

Using a systematic approach, the multiple clinical, genetic and environmental factors were logically analysed and reduced down to significantly associated factors using the need for surgical intervention as the primary outcome variable. All significantly associated factors from the genetic, clinical and environmental data analysis were then combined to produce a significantly associated single multi-factor model which identified the patient as being either at risk or with no risk for the need of surgical intervention. An example figure (actual factors and values not depicted) to show this methodological process has been depicted in Figure 4.2. Based on this diagnosis it was assumed that a treatment decision could be made by the clinical or gastroenterologists.

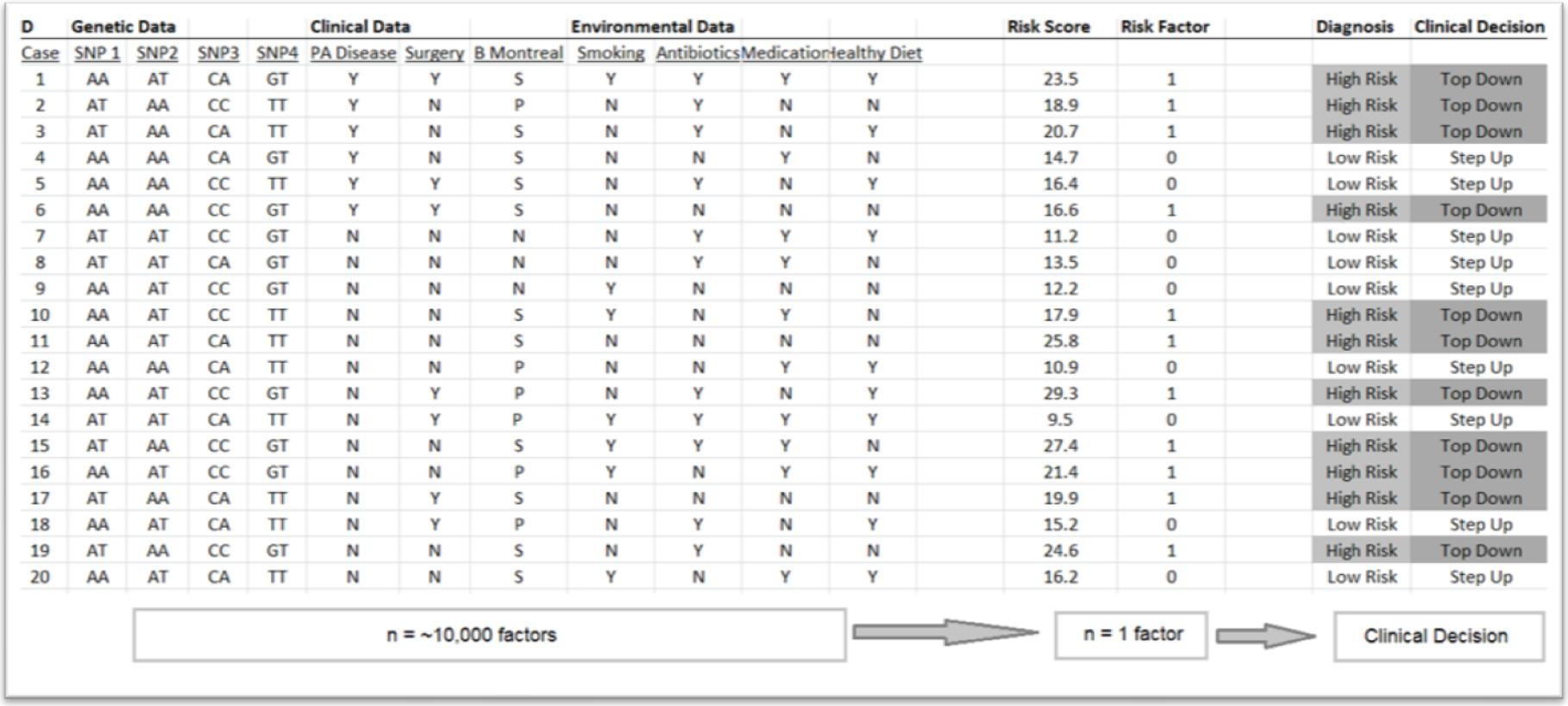

Figure 4.2 Systemic, Step-wise Data Reduction Analysis

Genotype data was cleaned and sorted to create files suitable for analysis in Plink and SPSS, and disease risk phenotype was incorporated into this file for further analyses. This data was examined for any discrepancies with original studies performed on the same cohort and no differences were found.

The systematic 5-step analysis strategy was performed as shown in Figure 4.3 and is subsequently explained.

Step One: Using the software package Plink, standard case/control analyses were performed to test for association between gene variants and the need for surgery as the clinical outcome (model option). This identified the genetic factors associated with the primary clinical outcome *(n = 9)*.

Step Two: Using SPSS, clinical factors were statistically analysed using unadjusted multi-factor logistic regression models. All clinical factors (*n=27*) were divided into two variable subgroups: (*i*) Disease History and (*ii*) Medical Treatment History. After a preliminary model analysis, adjustments using forward conditioning were performed as part of the logistic regression analysis and significant factors were entered into the model.

Step Three: All environmental factors (*n=28*) were divided into three groups of variables for this analysis: (*i*) Smoking exposure, (*ii*) Diet, and (*iii*) Household. SPSS was then used to perform an unadjusted multiple logistic regression model. After the preliminary model analysis, forward conditioning was performed by using the subsequently analysed factors that were entered into the model.

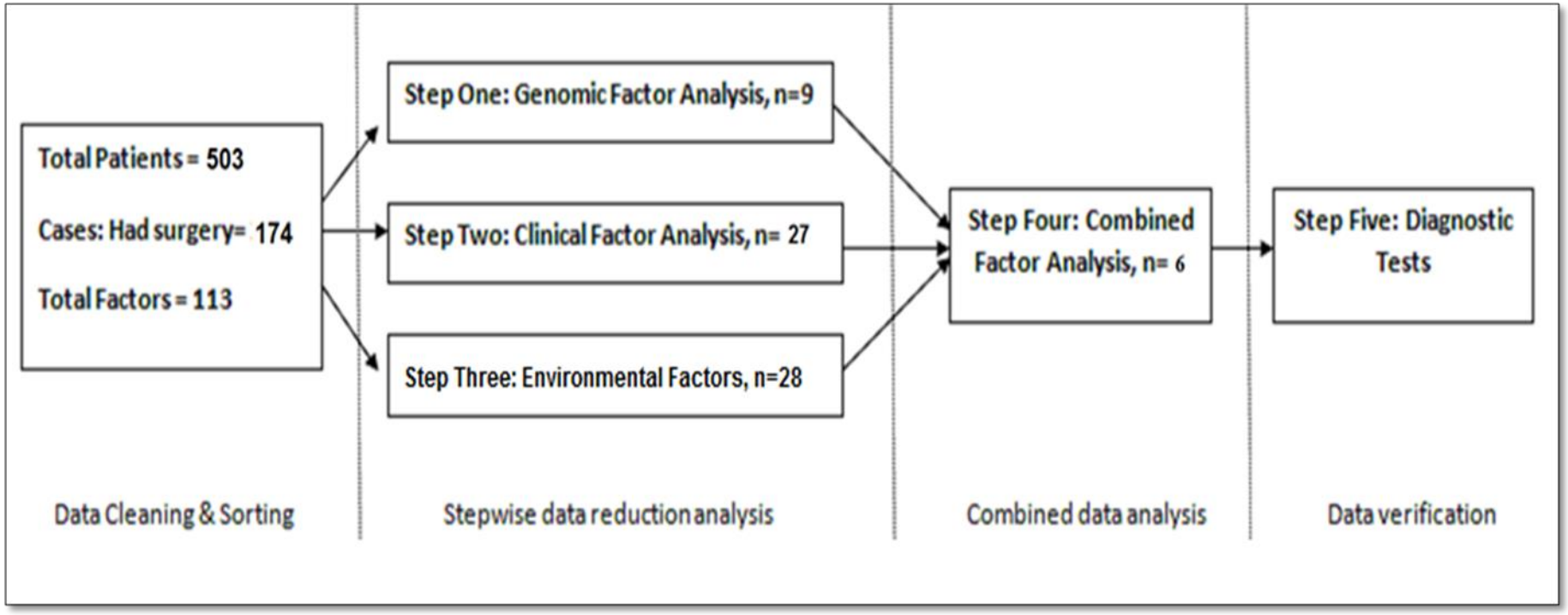

Figure 4.3 Stepwise Data Reduction Analysis Method

Step Four: This step involved a multi-factor analysis, model, by taking all of the previously significantly associated factors from all adjusted individual models that arose from analysis in previous steps, and entered into a forward conditioned multiple logistic regression model collectively. The multi-factor model was also adjusted for patient age and gender factors.

Step Five: After identification of significantly associated predicting factors of the need for surgery, stepwise statistical diagnostic tests were performed to verify the predictive probability and accuracy of the multi-factor envirogenomic risk profile model (Figure 4.3).

4.3.3.5 Prospective Analysis Results

Step One: Genomic Factor Analysis

Plink analysis of the gene variants considered separately revealed that the NOD2 genotype was the sole factor significantly associated with the need for surgery in this CD patient cohort (Table 4.3a). After running binary logistic regression using stepwise forward conditioning on all variants the NOD2 gene still remained as the most significantly associated risk factor for need for surgery in the cohort *(OR = 1.601, P = 1.6 x 10^{-3})* (Table 4.3b).

Table 4.3 Genomic Factor Analysis Results

a. Genomic factor analysis model using unadjusted multiple logistic regression (CHR=chromosome, SNP = single nucleotide polymorphism, MAF = minor allele frequency, CHISQ = chi square, P = P value)

CHR	SNP	Gene	Minor/Major Allele	MAF	Genotype Frequencies		Genotypic		Allelic	
					Case (n=174)	Control (n=333)	CHISQ	P	CHISQ	P
1	rs11209026	IL23R	A/G	0.030	0/10/155	0/25/287	-	-	0.582	0.446
1	rs35829419	NALP3	A/C	0.045	0/15/153	1/21/305	1.531	0.465	0.539	0.462
2	rs2241880	ATG16L1	C/T	0.419	31/79/58	56/162/109	0.302	0.859	0.000	0.984
4	rs6822844	IL21	T/G	0.314	11/83/73	27/180/124	1.915	0.384	1.512	0.219
5	rs13361189	IRGM	T/C	0.089	1/28/139	2/49/278	0.267	0.875	0.222	0.638
5	rs4958847	IRGM	G/A	0.142	1/46/122	4/70/254	2.494	0.287	1.076	0.2995
16	**NOD2***	**NOD2**	**T/A**	**0.347**	**18/77/68**	**16/128/168**	**9.360**	**0.009**	**8.510**	**0.004**
19	rs2043211	CARD8	T/A	0.278	13/67/87	32/136/162	0.680	0.712	0.645	0.424
22	rs4821544	NCF4	T/C	0.333	22/66/77	28/120/167	3.013	0.222	3.016	0.082

* 702W, 908R and 1007fs mutations

b. Genomic factor analysis model using adjusted step-wise forward conditional logistic regression (B = coefficient for the constant in the null model/intercept, S.E. = standard error, OR = odds ratio, C.I. = confidence interval)

Predictor	B	S.E.	P Value	OR	95% C.I.	
					Lower	Upper
NOD2	0.471	0.196	0.016	1.601	1.091	2.348

Step Two: Clinical Factor Analysis

The twenty-seven clinical factors were divided into two groups of variables: (***i***) Disease History and (***ii***) Medical Treatment History. The Disease History variables included variables relating to the history of any type of disease/illness for each patient. The Medical Treatment History variables included variables relating to the patients history of any type of medical treatment.

Disease History: The unadjusted logistic regression analysis model for the factors within this group is shown in Table 4.4a. Table 4.4b shows the adjusted multiple forward conditioning logistic regression analytical model. The final model shows that after stepwise data reduction analysis those patients who had ever had perianal disease *(OR = 2.695, P = 1.0 x 10^{-4})* were more likely to require surgery.

Medical Treatment History: The unadjusted logistic regression analysis model for the factors within this group is shown in Table 4.5. The final adjusted step wise forward conditioning regression model showed that there were no significantly associated variables from this group of factors.

Table 4.4 Results of Multi-Factor Logistic Regression for Disease History Clinical Factors

a. Unadjusted multiple logistic regression model for disease history clinical factors

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Perianal disease	1.022	0.215	0.000	2.779	1.822	4.238
Depression	0.426	0.252	0.091	1.531	0.934	2.510
Asthma	0.033	0.233	0.886	1.034	0.655	1.631
Eczema	0.455	0.257	0.077	1.576	0.952	2.608
Glandular Fever	0.269	0.286	0.346	1.309	0.747	2.293
Kidney Stones	0.055	0.742	0.940	1.057	0.247	4.525
Liver Disease	0.082	0.524	0.876	1.085	0.389	3.030
Mental illness	0.048	0.731	0.948	1.049	0.250	4.393
Bronchiectasis	0.150	1.470	0.919	1.161	0.065	20.708
Cancer	0.730	0.487	0.134	2.074	0.799	5.383
Tonsillectomy	0.060	0.224	0.790	1.061	0.685	1.645
Chole	0.032	0.408	0.938	1.032	0.464	2.296
Grommett	0.380	0.574	0.508	1.463	0.474	4.509

b. Adjusted step wise forward conditioning model for disease history clinical factors

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Perianal disease	0.991	0.211	0.000	2.695	1.783	4.074

Table 4.5 Results of Unadjusted Logistic Regression Analysis for the Medical Treatment History Clinical Factors

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Ever used immunomodulaters	1.029	0.595	0.084	2.797	0.871	8.980
Immunised against measles	0.185	0.974	0.849	1.203	0.178	8.123
Immunised against mumps	1.332	0.926	0.150	3.787	0.617	23.233
Immunised against TB*	1.151	0.707	0.104	3.163	0.791	12.649
Antibiotic consumption	0.062	0.852	0.942	1.063	0.200	5.644
Current OCP* consumption	0.498	1.355	0.713	1.645	0.116	23.426
Ex OCP* consumer	1.249	1.310	0.340	3.487	0.267	45.492
Antibiotic consumption during infancy	0.543	0.910	0.551	1.721	0.289	10.234
Antibiotic consumption during childhood	1.747	1.520	0.250	5.740	0.292	112.847
Antibiotic consumption during adolescence	2.532	1.756	0.149	12.577	0.402	393.102
Medication consumption during infancy	0.212	2.526	0.933	1.236	0.009	174.648
Medication consumption during childhood	2.128	2.224	0.339	8.400	0.107	657.071
Medication consumption during adolescence	0.655	1.064	0.538	1.925	0.239	15.505

*****(TB = Tuberculosis, OCP= Oral contraceptive pill)

Step Three: Environmental Factor Analysis

The 28 environmental factors were divided into three groups of variables: (***i***) Smoking, (***ii***) Diet and (***iii***) Household. Smoking variables included all variables related to smoking. Diet variables included all variables relating to the patients diet. Household variables included all variables relating to the patients household characteristics.

Smoking: There were 6 factors that were included in this group of variables that were analysed using unadjusted logistic regression analysis (Table 4.6a). In the next step, step wise forward conditioning logistic regression analysis was performed on these 6 factors and the model (Table 4.6b) shows that after the data reduction analysis those patients who were post diagnosis smokers *(OR = 5.359, P = 1.200 x 10^{-3})* or ex-smokers at diagnosis *(OR = 2.706, P=1.000 x 10^{-4})* remain as significantly associated with the need for surgery.

Household: A total of 16 factors were analysed within this group of variables. After unadjusted logistic regression models and adjusted forward conditioning regression models were created no factors showed any significant association with the need for surgery (Table 4.7).

Diet: This group of variables had 6 factors and after unadjusted regression and a step wise forward conditioning data reduction; no factors were seen as significantly associated with the need for surgery (Table 4.8).

Table 4.6 Results of Multi-factor Logistic Regression for Smoking Variables.

a. Unadjusted logistic regression for smoking environmental factor analysis

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Mother smoker	0.036	0.239	0.880	1.037	0.649	1.655
Ever smoker	0.222	0.235	0.346	1.248	0.787	1.980
Smoker at dx*	1.539	0.684	**0.024**	4.661	1.220	17.809
Post dx* smoker	1.984	0.673	**0.003**	7.723	1.946	27.173
Never smoker	0.260	0.194	0.181	1.297	0.886	1.897
Ex-smoker at dx*	2.677	0.704	**0.000**	14.536	3.660	57.731

* dx: diagnosis

b. Adjusted step wise forward conditioning logistic regression model for smoking environmental factors

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Post dx* smoker	1.679	0.668	0.012	5.359	1.448	19.834
Ex-smoker at dx	0.995	0.272	0.000	2.706	1.588	4.611

*Post dx: Post diagnosis, dx: diagnosis

Table 4.7 Multi-factor regression Results for Household Variables

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Public Swimming	0.059	0.170	0.730	1.061	0.760	1.480
Pool Infant	0.576	0.355	0.104	1.778	0.888	3.564
Pool Child	0.242	0.254	0.341	1.273	0.774	2.094
Pool Adol*	0.120	0.246	0.626	1.127	0.696	1.826
Sand pit Infant	0.060	0.168	0.720	1.062	0.764	1.475
Sand pit Child	0.023	0.176	0.895	1.024	0.725	1.446
Sand pit Adol*	-0.412	0.231	0.074	0.663	0.422	1.041
Farm Infant	0.113	0.206	0.583	1.120	0.748	1.676
Farm Child	0.182	0.198	0.358	1.199	0.814	1.766
Farm Adol*	0.200	0.210	0.339	1.222	0.810	1.843
Smoke Infant	-0.321	0.171	0.060	0.725	0.519	1.014
Smoke Child	-0.233	0.166	0.161	0.792	0.572	1.097
Smoke Adol*	-0.234	0.165	0.156	0.791	0.573	1.093
Share bedroom Infant	-0.096	0.166	0.564	0.909	0.656	1.258
Share bedroom Child	0.053	0.164	0.748	1.054	0.764	1.454
Share bedroom Adol*	0.000	0.167	0.999	1.000	0.720	1.388

*Adol = adolescence

Table 4.8 Multi-Factor Regression Results for Diet Variables

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
Breastfed	0.104	0.218	0.634	1.109	0.724	1.700
Alcohol	0.128	0.211	0.546	1.136	0.751	1.719
Vegetarian	0.231	0.454	0.611	1.260	0.517	3.068
Takeaways Infant	0.157	0.344	0.648	1.170	0.596	2.297
Takeaways Child	0.036	0.283	0.899	1.037	0.595	1.805
Takeaways Adol*	0.235	0.240	0.326	1.265	0.791	2.024

* Adol: Adolescence

Step Four: Multi- Factor Analysis

After analysing all of the genetic, clinical and environmental factors independently, the next step involved taking all of the significantly associated variables from each of these three factor groups and analysing them collectively. The factors NOD2 gene, Perianal disease, ex-smoker at diagnosis and post-diagnosis smoker stood out as significant and all these factors underwent a forward conditioning logistic regression multi-factor model analysis (Table 4.9).

The next step involved using our combined multi-factor model and adjusting the results for age and gender. The final, adjusted multi-factor model showed that current age *(OR=1.012, P=4.40 x 10^{-3})* was significantly associated with the need for surgery along with the four previously identified factors which included the NOD2 gene *(OR=1.607, P=2.30x10^{-5})*, Perianal disease *(OR=2.847, P=4.0x10^{-6})*, ex-smoker at diagnosis *(OR=2.405, P=1.10x10^{-3})* and post-diagnosis smoker *(OR=6.312, P=7.40x10^{-3})* (Table 4.10). Figure 4.4 shows the final regression equation that can be derived from the final multi-factor envirogenomic model.

Table 4.9 Results of Multi-Factor Forward Logistic Regression Analysis

Predictor	B	S.E.	P value	OR	95% C.I.	
					Lower	Upper
NOD2	0.469	0.207	0.024	1.599	1.065	2.400
Perianal disease	0.971	0.222	0.000	2.641	1.708	4.082
Ex-Smoker at dx*	0.782	0.263	0.003	2.187	1.307	3.659
Post dx* Smoker	1.705	0.682	0.012	5.500	1.445	20.941

*dx: diagnosis

Table 4.10 Results of Multi-Factor Forward Logistic Regression Analysis, Adjusted for Age and Gender

Predictor	B	S.E.	P Value	OR	95% C.I.	
					Lower	Upper
NOD2	0.474	0.208	0.023	1.607	1.068	2.417
Perianal disease	1.046	0.227	0.000	2.847	1.823	4.447
Ex-smoker at dx*	0.878	0.268	0.001	2.405	1.422	4.069
Post dx* smoker	1.842	0.687	0.007	6.312	1.640	24.285
Current Age	0.012	0.006	0.044	1.012	1.000	1.024

*dx: diagnosis

Logit P of Crohn's: *-2.457 + 0.474 x NOD2 + 1.046 x Perianal Disease + 0.878 x Ex-Smoker at Diagnosis + 1.842 x Post-Diagnosis Smoker + 0.012 x Current Age*

Figure 4.4 Logistic Regression Equation for the Combined Multi-Factor Model

Step Five: Diagnostic Testing

To test the predictive value of the final combined model derived from the previous analytical steps, various investigative tests were performed on the individual probabilities generated from the regression model. First, ROC curves displaying the AUC were created for the combined multi-factor model as well as for models of the clinical, environmental and genetic

factors considered individually (Figure 4.5). As an individual factor, the genomic model produced an AUC of 56% *($P = 3.70 \times 10^{-3}$)*, the clinical model produced an AUC of 60% *($P = 1.0 \times 10^{-4}$)* and environmental model produced an AUC of 59% *($P = 1.0 \times 10^{-4}$)*.

When genomic and clinical factors are combined the AUC slightly increases to 63% *($P = 1.0 \times 10^{-4}$)*, when genomic and environmental factors are combined, the AUC also slightly increases to 63% *($P = 1.0 \times 10^{-4}$)*, and when clinical and environmental factors are combined the AUC also increases slightly to 65% *($P = 1.0 \times 10^{-4}$)*.

Whilst the individual factor models offered some predictive value it was found that the multi-factor envirogenomic model produced the highest predictive value of 68% *($P = 1.0 \times 10^{-4}$)* (Table 4.11), and hence this model was used for further diagnostic testing.

In order to dichotomize our curve, and hence stratify the patients into either a high risk or a low risk category, an optimal cut off point needed to be chosen.

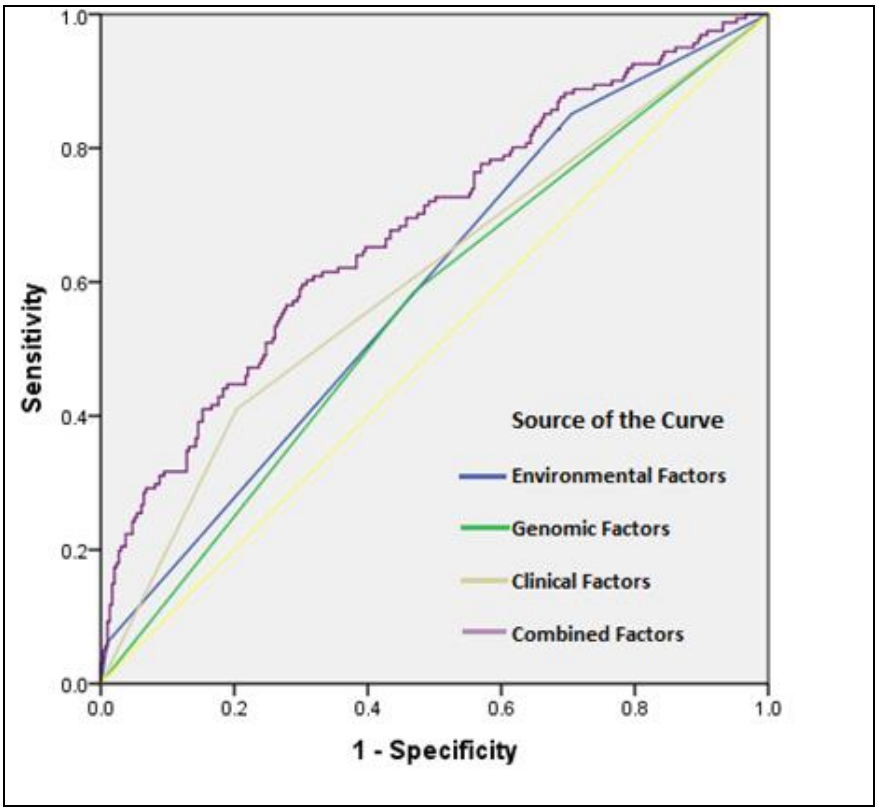

Figure 4.5 Multi-Factor ROC Analysis Results

Table 4.11 Individual and Multi-Factor AUC Results

Factors	AUC	S.E.	P value	95% C.I.	
				Lower	Upper
Genomic	0.558	0.028	0.037	0.504	0.613
Clinical	0.603	0.028	0.000	0.548	0.659
Environmental	0.590	0.027	0.001	0.537	0.644
Genomic + Clinical	0.630	0.028	0.000	0.576	0.684
Genomic + Environmental	0.629	0.027	0.000	0.576	0.682
Clinical + Environmental	0.654	0.026	0.000	0.604	0.705
Genomic + Clinical + Environmental	0.681	0.027	0.000	0.629	0.733

To do this, two points were selected at 80% sensitivity and 80% 1-specificity (Figure 4.6). Diagnostic calculations were performed on both cut off points (Table 4.12) and it was found that the second cut off point at 80%

1-specificity showed the higher OR of 3.169 *(P = 1.0 x 10^{-4})*, higher PPV of 0.545 and also had a higher attributable risk of 0.271.

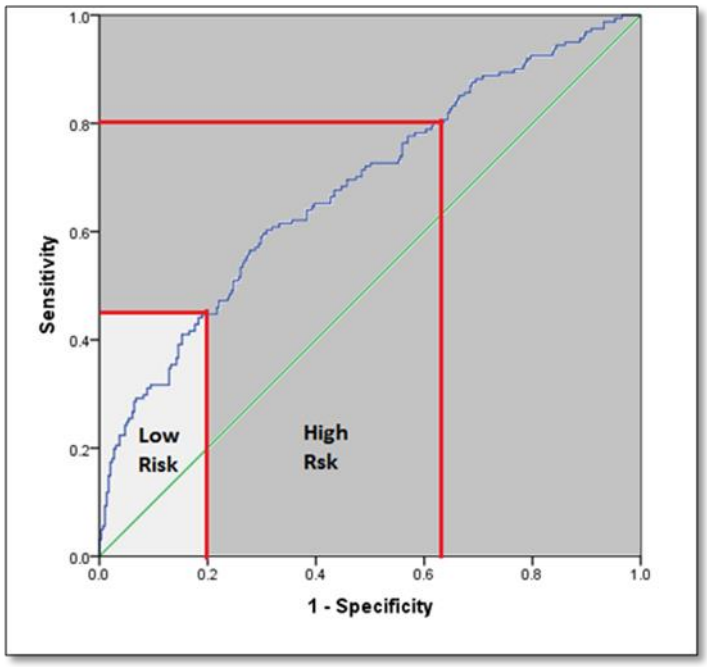

Figure 4.6 Cut-off Points Selection

Table 4.12 Diagnostic Calculations on Selected Cut-off Points

	OR	P value	Sens	Spec	PPV	NPV	AR	RR
Cut off 1 *	2.397	0.000	0.801	0.373	0.411	0.775	0.185	1.823
Cut off 2*	3.169	0.000	0.447	0.797	0.545	0.725	0.271	1.986

*Cut off 1 = 80% Sensitivity, Cut off 2= 80% 1-Specificity

4.3.3.6 Prospective Analysis Discussion

Many studies have shown that clinical and environmental factors are essential components of the pathogenesis of CD and are primarily responsible for its growing incidence around the globe (Gearry *et al*, 2006, Siegel *et al*, 2011, Lewis *et al*, 2007, Manolio *et al*, 2009, Gearry *et al*, 2010). Various research analyses have investigated the environmental case for CD when there is a strong link between genetics and CD (Kugathasan *et al*, 2006).

The natural disease course for CD is highly variable among patients ranging from mild cases requiring dietary modification to severe cases requiring surgery. Despite the fact that direct epidemiological evidence for the presence of potential gene–environment interactions in IBD aetiology is limited, other lines of evidence point to the presence of such interactions. For example, the heterogeneity in risks from environmental exposures (e.g. diet, hygiene) across different populations may be related to the presence of gene–environment interactions (Kugathasan *et al*, 2006). Therefore, clinicians need to be accurate in their diagnosis and prediction of disease progression when treating patients in order to personalise medical therapy and optimise outcomes. This involves not only assessing genetic risk factors but also the growing number of clinical and environmental risk factors associated with disease progression and outcome.

Early intervention and diagnosis is necessary for patients so that optimal treatment can be achieved and progressive disease symptoms can be avoided. One goal of clinicians is to utilise predictive factors to accurately categorise patients as either high or low risk for requiring surgery. Hence, clinicians could use such algorithms to select the optimal treatment approaches and deliver personalised medicine. This could include top-down therapy for those with higher risk and a more traditional approach for those with a lower risk of requiring subsequent surgery.

The stepwise data reduction method applied to this population-based IBD cohort has identified genetic, clinical and environmental factors that can be combined to predict the need for IBD-related surgery in CD patients. The multi-factor envirogenomic model identified from this study included the NOD2 genotype, perianal disease, smoking, and age factors. This combined factor model yielded an increased predictive value from when each factor was considered individually.

The NOD2 gene variants that were analysed in this study have been previously associated with the risk of developing complicated CD. Smoking has also been associated with the development and progression of CD increasing the risk for IBD-related surgery (Roberts *et al*, 2010, Gearry *et al*, 2010, Gearry *et al*, 2006, Nasir *et al*, 2012, Lewis *et al*, 2007). Similarly, perianal disease is associated with an increased risk of progressive complicated CD (Nasir *et al*, 2012, Lakatos *et al*, 2009, Tarant *et al*, 2008, Yang *et al*, 2011) and resectional surgery.

Others have also demonstrated the clinical utility of combining clinical, serological and genetic CD data to predict the progression and severity of CD (Weersma *et al*, 2009). It has also been previously shown that the risk of surgery is associated with combined genetic and clinical factors – NOD2 genotype and perianal disease (Nasir *et al*, 2012). However, studies where clinical, genetic and environmental data have been combined and analysed in a systematic way to yield a combined envirogenomic score for determining the risk of complicated disease and surgical intervention has not been performed previously.

Previous studies have also shown independent predictor models for various individual risk factors (Wei *et al*, 2011) as well as formulated tools to predict the development of complicated CD using independent risk factors (Siegel *et al*, 2011). However the multi-factor model developed in this study produces a much higher predictive probability for the risk of surgery then when any of the risk factors are considered independently, indicating the usefulness of the method of research utilised in this study. Consequently, our multi-factor envirogenomic risk model has taken a step closer to identifying patients at risk for the need of surgery potentially allowing early personalized therapeutic decisions to be made that can prevent the development of complicated disease.

From this study it can be concluded that CD patients from this New Zealand population who possess these clinical, genetic and environmental disease factors are at higher risk of IBD-related surgery. It has been shown in a previous study using this population that IBD in Canterbury is as common as other Western countries and that CD incidence and prevalence is among the highest ever reported in this population (Gearry *et al*, 2006). Consequently, it can be implied that similar demographically corresponding populations would have comparable outcomes. Further analyses of other large and well-characterised cohorts, including prospective assessment, may confirm these findings and it is also possible that other markers of disease severity and prognosis may be able to be integrated into this model. This and related research highlights the clinical usefulness of genomic, epidemiological and clinical research in predicting outcomes for patients with CD when performed using a combined factor analysis approach.

4.3.4 Study Limitations

Possible limitations may occur during the statistical analytical steps during the analysis of the data, including performing ROC examinations. One of the major limitations of ROC analysis is that data must be divided into two states. This raises another problem of whether the data will clearly fall into one state or the other. In the case of presence or absence of a disease, this is already divided into two states. The sample size must also be large enough for the effects to be real and significant. It should also be representative, if possible, of the actual population. The sample should include the entire spectrum of each state of diagnostic truth. The major drawback to threshold-based approaches is that they often lack the sensitivity and specificity needed for accurate classification.

As a result of performing diagnostic testing on our data we have tried to overcome this problem within this analysis. Another problem is the control of other parameters which could affect the diagnosis of the observer. Implications that may occur practically can include possible problems that can include data analysis and statistical mistakes in calculations that can lead to false results. Also, managing time wisely and effectively can become a problem if not conducted properly due to the expanse and time consuming nature of this project.

Due to the limited genotypic, clinical and environmental data currently available in the data cohort, the results obtained might not be indicative of being entirely applicable to the general population as a whole. The main goal of this analysis was to carry out a 'comprehensive statistical method' that would be able to provide some indication of significant results. In future, using large-scale and more appropriate data it would be feasible to assume that better significance and power estimates could be achieved. It is hoped that by performing replication analysis on varying data cohorts from around the world, the possibly limiting criterion witnessed will be overcome and the future obtained results will be applicable to the entire general population.

4.3.5 Ethical Issues

Possible ethical issues within this section of the project have already been covered and taken care of before the commencement of the Canterbury IBD project by data collectors and organizers. Relevant authorizations and clearance procedures were requested and granted prior to the gathering of the data from participants for this project by principal supervisors of the Canterbury IBD project. Patients were kept anonymous throughout the data collection and analysis procedures and the

information collected was used solely for the purposes of research. Patient information was kept under strict security controls in secure databases.

4.4 The Queensland Institute of Medical Research Replication Study

In order to verify that the initial envirogenomic risk profiling method was a worthwhile approach to identify risk factors associated with IBD-related surgery, it was necessary to replicate the method using another independent IBD data set. This replication took place with data provided from collaborators at the Queensland Institute of Medical Research (QIMR) in Brisbane, Australia. Data that was available from the Canterbury IBD cohort was also included as part of this investigation as a separate cohort.

Two primary factors associated with development of complicated CD are cigarette smoking before or at diagnosis (Lakatos *et al*, 2010, Gearry *et al*, 2010) and perianal disease (Lakatos *et al*, 2010, Tarant *et al*, 2008, Yang *et al*, 2011, Nasir *et al*, 2012). Patients exposed to these factors may be more likely to develop complicated disease behaviour and, in many cases, surgical intervention will be required, often with the risk of disease recurrence post-surgery (Bernell *et al*, 2000). This subgroup of severe CD patients might, therefore, benefit from intensive early treatment with biological therapies – the so-called top-down approach to treatment (Baert *et al*, 2007, Bouguen *et al*, 2011). However, to treat all CD patients in this fashion is not economically or medical sensible and as such there is a need for clinicians to be able to more reliably predict the development of complicated disease so that early intensive management of the disease can be provided to the patients most likely to benefit from it.

The aim of this study was to determine the association of two primary factors (Perianal disease and Smoking) for which data was readily available from the QIMR cohort, in order to identify a multi-factor model that predicts the risk of surgery and can be interpreted at the clinical level so that patients can benefit from early disease treatment and avoid the need for unnecessary surgical interventions. Only data that was uniform and complete for each of these variables was incorporated for this research study.

4.4.1 Data Analyses Methods

***Patient selection and Data Ascertainment*:**

Data was requested from medical centres with a specialized interest in IBD across Australia and New Zealand. Incorporated clinical data for this investigation was collected from CD cohorts across five IBD treatment centres from Australia and New Zealand - Queensland Institute of Medical Research (QIMR), Bancroft Centre (Brisbane, Queensland), Flinders Medical Centre (FMC) (Bedford Park, South Australia), Royal Adelaide Hospital (Adelaide, South Australia), Fremantle Hospital (Perth, Western Australia) and the Christchurch Hospital, Canterbury IBD cohort (Christchurch, New Zealand) (Table 4.13). The inclusion of the Canterbury IBD cohort was decided upon, so as to increase the amount of data available and allow for a more comprehensive analysis of Australian and New Zealand CD patients, covering a broader regional area. After discussion with the associated data collectors and gastroenterologists, it was concluded that the inclusion of the Canterbury IBD cohort would not cause any concern for possible bias based on the statistical methods applied to the data. Moreover, the QIMR study was not a strict “replication” study with complete independence from the

Canterbury study. The QIMR study utilized a different design which specifically allowed assessment of clinical and environmental factors on IBD related surgery (no genetic data was available for this cohort).

The study design was a retrospective association analysis of 2,262 patients diagnosed with CD by the respective Centre's gastroenterologists. All patients were formally diagnosed using the Lennard-Jones criteria for CD patients across all cohorts (Lennard-Jones *et al,* 1989). All patients were racially from similar backgrounds and of Caucasian ethnicity, and therefore not a cause for any possible limitations. Those patients with missing or incomplete information for any of the variables that were required for this specific analysis were excluded from this data set. After data cleaning and setup, a combined total of 1,725 patients were available for analysis.

Table 4.13 QIMR Replication Study Patient Characteristics

Cohort	Patients (%[1])	Male/Female(%[2])	Any CD Surgery (%[2])	PD[3] (%[2])	Smoker[4] (%[2])	PD+ Smoker (%[2])
Brisbane	655 (38.0)	282(43.1)/373(56.9)	470 (71.8)	203 (37.0)	322 (49.8)	89 (13.6)
Adelaide	172 (10.0)	76(44.4)/96(55.6)	102(59.3)	64(37.2)	11 (7.0)	5 (2.9)
FMC	395 (22.9)	172(43.5)/223(56.5)	260 (65.8)	97(35.5)	140(95.2)	37(9.4)
Canterbury	503 (29.2)	186(37.0)/317(63.0)	174(34.6)	137(27.2)	245(50.4)	59(11.7)
Combined	1725 (100)	716 (41.6)/1009(58.5)	1006 (58.3)	501 (33.5)	718 (49.9)	190 (11)

[1]:% total, [2]: %total within cohort, [3]: PD: Perianal disease, [4] Smoker = Ever smoker before/at diagnosis

***Predictor variables*:**

The predictor variables that were assessed were if the patients (***a***) "ever suffered perianal disease" or (***b***) "ever smoked before or at CD diagnosis (Smoking)". These variables were selected for the study because they have been shown to be associated with CD risk and disease complication in a number of studies (as highlighted before) (Krishnaprasad *et al*, 2012). All IBD clinicians from the consortium agreed that the predictor variables were ascertained in a consistent fashion across the different centres and that there were no discrepancies in the format for each of these variables.

***Clinical Outcome*:**

The need for CD-related surgery was considered as the primary clinical outcome for this study incorporating any type of intestinal resection, and any type of non-resection surgery such as drainage of perianal abscesses was not included. This outcome was defined by gastroenterologists who confirmed the occurrence of any type of intestinal bowel resection as a direct result of CD at the time of recruitment on patient records. Patients were categorized and coded dichotomously as follows:

0: have had no CD-related surgery at the time of recruitment, or

1: have had at least one CD-related surgery at the time of recruitment.

***Statistical Analyses*:**

Patients with missing data for the primary outcome variable or the predictor variables analysed in this study were removed, leaving data for 1792 patients for analysis. Multi-factorial logistic regression was performed by

considering the need for CD-related surgery as the primary clinical outcome. Stepwise forward conditioning was performed to create regression models that showed the selected significant risk factors associated with the need for surgery. Using the resulting probabilities from the regression models, a ROC analysis was performed for the individual factors, as well as the multi-factor model to test the value of a combined factor analysis. Diagnostic testing was also performed on the final multi-factor model to calculate the predictive probability and accuracy of the combined model.

4.4.2 Results

An individual logistic regression analysis was first performed to check if the two chosen predictive factors were in fact significant in the individual cohorts selected for this transnational study. Results showed that Perianal disease was significant in all cohorts (Table 4.14). The Brisbane cohort was significant for both smoking and Perianal disease variables.

Table 4.14 Individual Cohort Regression Results

Cohort	Variable	B	S.E.	P Value	OR	95% C.I. Lower	95% C.I Upper
Brisbane	PD	1.17	0.24	0.00	3.21	2.01	5.12
Brisbane	PD&CIG	0.68	0.21	0.00	1.97	1.31	2.95
Flinders Medical Centre	PD	1.08	0.45	0.02	2.95	1.23	7.07
Royal Adelaide Hospital	PD	1.96	0.45	0.00	7.13	2.93	17.37
Fremantle Hospital	PD	0.44	0.22	0.05	1.56	1.01	2.40
Canterbury	PD	1.01	0.21	0.00	2.75	1.83	4.12

PD = perianal disease, CIG = smokers

The Multi-factor logistic regression analysis of the two predictor variables showed that both Perianal disease *(OR = 3.14, P = 1.00 x 10^{-6})*, and Smoking *(OR = 6.03, P = 1.00 x 10^{-6})* were associated with an increased risk of surgery (Table 4.15).

Table 4.15 Multi-Factor Logistic Regression Analysis Results

					95% C.I.	
Variable	**B**	**S.E.**	**P Value**	**OR**	**Lower**	**Upper**
Perianal Disease	1.15	0.14	1×10^{-6}	3.14	2.41	4.11
Smoking	1.80	0.13	1×10^{-6}	6.03	4.67	7.80

In order to better understand and comprehend the results obtained, each cohort was analysed independently for these two associated variables. Statistics for the smoking and perianal disease variables were analysed for each cohort and are tabulated in Table 4.16 below.

Table 4.16 Independent Cohort Analysis for Significantly Associated Variables

					95% C.I.	
Cohort/Variable	**B**	**S.E.**	**P Value**	**OR**	**Lower**	**Upper**
Brisbane						
PD	1.165	0.239	0.000	3.205	2.005	5.122
Smoker	0.676	0.207	0.001	1.967	1.312	2.950
Flinders Medical Centre						
PD	1.082	0.446	0.015	2.952	1.233	7.068
Royal Adelaide Hospital						
PD	1.964	0.454	0.000	7.127	2.925	17.365
Fremantle Hospital						
PD	0.441	0.222	0.047	1.555	1.006	2.403
Canterbury						
PD	1.010	0.207	0.000	2.747	1.832	4.119

PD = perianal disease, CIG = smokers

ROC analysis was then performed to determine the predictive accuracy of the regression probabilities by creating an area under the curve (AUC) for the multi-factor model, as well as for both the individual factor models for comparison (Figure 4.7).

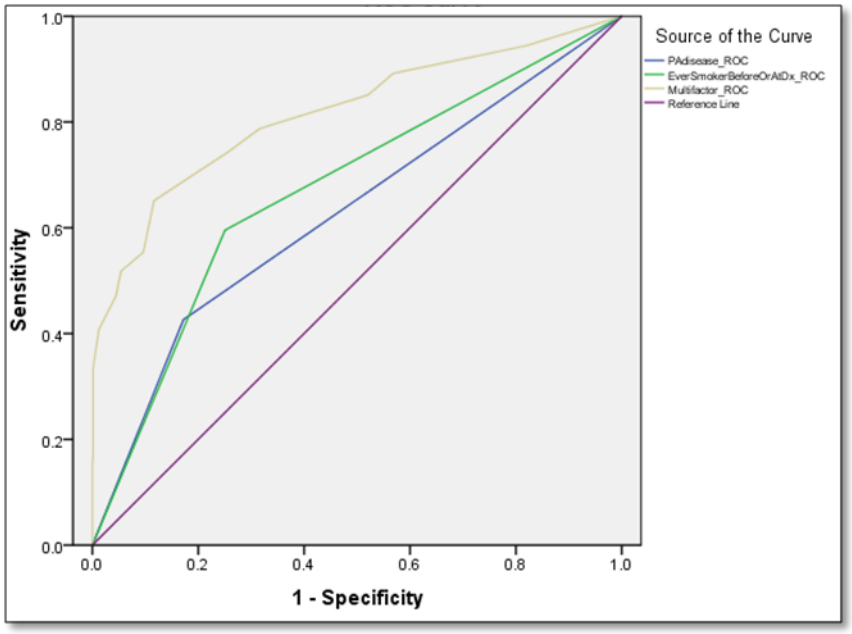

Figure 4.7 The AUC Analysis for the Combined Multi-Factor Model as well as the Individual Factors

Table 4.17 AUC Results for Individual Factors and Multi-Factor Model

				95% C.I.	
ROC Model	**AUC**	**S.E.**	**P Value**	**Lower**	**Upper**
Perianal disease	0.63	0.01	1×10^{-6}	0.60	0.65
Smoking	0.67	0.01	1×10^{-6}	0.65	0.70
Multi-factor Model	0.82	0.01	1×10^{-6}	0.80	0.84

* Model adjusted for cohort heterogeneity

Results showed that the multi-factor model produced the highest AUC of 0.82 *($P = 1.00 \times 10^{-6}$)* as compared to 0.63 for Perianal disease *($P = 1.00 \times 10^{-6}$)* and 0.67 for Smoking *($P = 1.00 \times 10^{-6}$)* (Table 4.17). This indicates that the multi-factor model offers substantially increased predictive value over either of the factors when considered independently. This multi-factor model suggests that a randomly selected patient from this multi-cohort dataset would have an 82% risk

of needing CD-related surgery if they possessed both predictor variables compared to those patients who possess neither.

Table 4.18 Cut-off Point Selection Diagnostic Calculations

	OR	P Value	Sensitivity	Specificity	PPV	NPV	MR	RR
Cut Off 1*	19.01	1×10^{-6}	0.93	0.59	0.47	0.96	0.32	10.55
Cut Off 2*	3.72	1×10^{-6}	0.59	0.72	0.94	0.18	0.39	1.15

*Cut off 1 = 80% Sensitivity, Cut off 2 = 80% 1-Specificity,

Using the multi-factor model to determine the best point of division to classify patients into diagnostic risk categories, further diagnostic calculations were performed on two selected cut-off points: (***i***) 80% 1-specificity and (***ii***) 80% sensitivity (Table 4.18). Results indicated that the second cut-off point at 80% 1-specificity had the higher PPV of 0.94 and the lowest misclassification rate of 0.39. This point was then used to dichotomize the multi-factor model in order to classify patients at a high or low risk towards the need for surgical intervention (Figure 4.8).

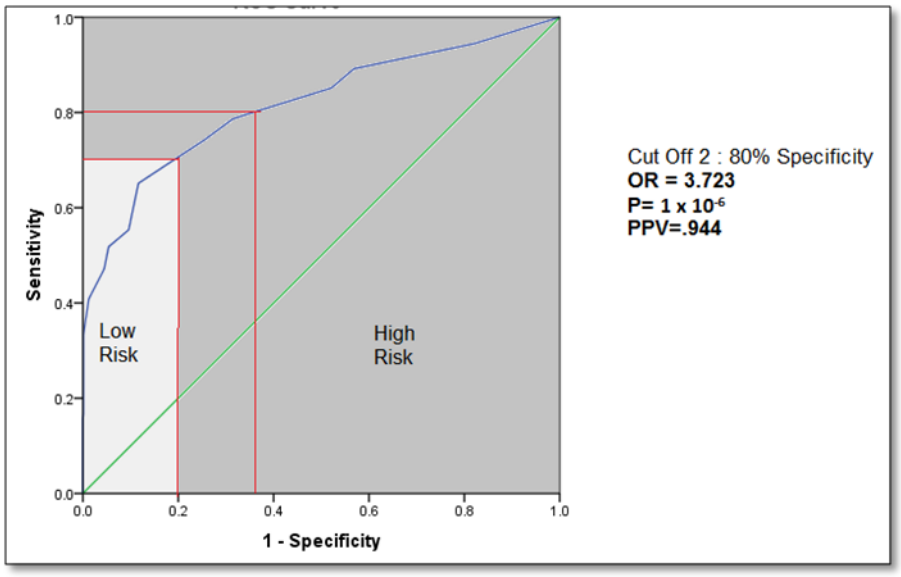

Figure 4.8 Cut-off Points Selection and Diagnostic Risk Division

This important diagnosis which can be derived from dichotomising the patients into a high or low risk classification, can be used by clinicians, gastroenterologists or medical practitioners to prescribe individual patients with the most appropriate and useful form of medical treatment available for them, thus fulfilling the goals of personalized treatment approaches.

4.4.3 Discussion

This large transnational investigation analysed two common clinical risk factors that have been previously associated with the need for surgical intervention in CD patients. This investigation has developed a model whereby patients can be stratified as either high or low risk towards the need for surgery. With a 94% PPV and only a 39% misclassification rate, considerably accurate diagnosis has been achieved based on only investigating two pieces of routinely available clinical data.

The impact of the patients smoking status before or at the time of diagnosis showed the highest correlation on the risk of CD surgery resulting in the strongest association shown in the combined model *(OR = 6.03, P = 1.0 x 10^{-6})*. Due to limitations of data available, the effect of time spent smoking during these durations was incomplete and in some cases unavailable and therefore cannot be compared as part of this investigation. Nevertheless, the study has shown that patients who smoked before or at the time of diagnosis are at a substantial risk of requiring CD surgery during their disease life course. Patients who had ever had perianal disease were also significantly more likely to require surgery *(OR = 3.14, P = 1.0 x 10^{-6})*.

Diagnostic testing revealed significant predictive probabilities for the multi-factor model as well. These predictive probabilities suggests that if patients from this cohort have perianal disease and were ever smokers before or at the time of diagnosis, then they are at an 82% higher risk of requiring CD-related surgery compared to those patients who were tested negative for the predictor variables analysed. This combined model approach at investigating disease risk predictors has shown that a significantly accurate diagnosis for the risk of CD surgery can be derived for those patients who possess these two factors.

After the dichotomization of the multi-factor model AUC to identify high and low risk patients, the second-cut off point proved to be the most valuable. A total of 1,674 cases were used to determine the diagnostic statistics of both cut-off points, with those cases with missing baseline parameters removed by SPSS. The second cut-off point identified 882 (49%) cases that were correctly diagnosed as being at risk for the need for surgery, compared to only 440 (24%) cases from the first cut-off point. A total of 606 (34%) patients had not had any CD-related surgery but were at high-risk for the need for surgery according to the combined multi-factor model second cut-off point, in comparison to 33 (2%) from the first cut-off point. This second-cut off point also had only 52 (2%) cases that had surgery but were not identified as at a high risk for the need for surgery according to the multi-factor model, however the first cut-off point had 494 (28%) that were not classified as high risk according to the multi-factor model but in actuality had previous CD-related surgery. The second-cut off point had a very high precision rate of 0.94 and only a 0.39 misclassification rate. Thus, the second-cut off point was chosen to dichotomize patients at being either high or low risk towards the need for surgery.

Based on this diagnosis, clinicians, gastroenterologists and/or medical practitioners can at least now more confidently prescribe individual personalized treatment regimens for their patients. It was also seen from this replicative analysis, that the initial Canterbury IBD data group analysis was noteworthy, and when a larger transnational data cohort was used, a higher predictive probability was obtained.

This type of information can be used to create diagnostic tools that can aid clinicians with prognostic analysis in an attempt to personalize the medical treatment therapies available for high risk patients and reduce the need for unnecessary surgical intervention. These promising results require further analysis using more data from other cohorts to improve the prognostic value of these results in future.

A number of studies have shown that early management of CD with optimal therapies can reduce hospital admissions, and potentially avoidable surgical interventions (Bernell *et al*, 2000, Vermeire *et al* 2007, Gearry *et al*, 2010). The course of the disease can also be influenced by the initiation of aggressive early treatment. It is, therefore, important to identify patients that may be at an increased risk of developing complicated disease behaviour as these patients have the most to gain from an early aggressive approach.

However, because of the many different clinical and environmental characteristics of CD, patients often respond differently to the medications that are currently available for the treatment of CD, and are usually prescribed generalised treatments based only on their known symptoms. It is, therefore, challenging for clinicians and/or gastroenterologists to decide and select the most effective type of

drug treatment that can provide each individual patient with the best medical therapy available for the best management of the disease. This signifies the importance and necessity for improved prognostic tests that can reliably predict disease progression pathways for CD.

Previous studies have analysed various CD predictive factors one at a time to assess their association with the development of complicated CD (Lewis *et al*, 2007, Siegel *et al*, 2011, Dubinsky *et al*, 2011). Studies have also looked at assessing the risk of intestinal resection for CD patients (Bouguen *et al*, 2011). This study however, investigated two key CD predictive factors when acting in *combination*, from a combined large transnational cohort analysis to assess the primary outcome for the need of CD-related surgery. By using a joint multi-factor model, we have shown that a much higher predictive probability can be achieved to show an association towards the risk of surgery.

The current Australian and New Zealand healthcare systems are not sufficiently flexible to enable the practice of personalized medical treatments for CD patients. Generally clinicians tend to start aggressive treatment therapies when complicated disease has already progressed because they tend to avoid the risk of over treating patients, even though treatments are readily available, due to the lack of predictive information currently known about CD. This type of research should hopefully provide useful guidance for clinicians to confidently provide appropriate personalized treatment to patients with reasonable justification.

Further investigations should include an increased number of clinical, environmental as well as genetic risk factors that can be possible predictors of disease and be analysed in individual CD cohorts. Eventually, it is hoped that by

analysing further CD risk predictors, researchers will be able to provide personalized medicine which may optimize the use of more expensive forms of medical therapy, reduce hospitalizations, and unnecessary surgery. Such an approach would be likely to lead to improved clinical outcomes and reduced costs for the healthcare system.

4.4.4 Ethical Issues

All study participants gave their written informed consent to their respective clinicians/gastroenterologists at the time of data collection and each database has the necessary local Human Research Ethics consent. All data sent for central analysis was kept anonymous to maintain patient confidentiality throughout the course of this research. All patient data was kept in secure databases on secure, private computers.

4.5 The Development of a Web Based CD Risk Calculator

This section of the research project aimed to design a computational tool which would provide a probabilistic prediction of a patient's risk of having CD based on their provided genetic, environmental and clinical data.

This computational tool was developed on the results obtained from the Canterbury IBD prospective study by helpful collaborators who set up this prototype risk calculator. Using the results obtained from the Canterbury IBD data prospective section of this project, a logistic regression equation identifying the key factors identified as associated with the risk of surgery was created. By incorporating this Logit formula, our expert IT collaborators were able to create a web-based clinical risk calculator tool.

The objective was to develop a user friendly computational tool that could be used in a clinical/medical environment by any clinician, gastroenterologist or medical practitioner at the time of diagnosis. When patient data is collected and a diagnosis for CD is determined, it was hoped that by inserting patient information into this calculator, based on the predictive disease risk factors a CD risk factor assessment could be determined for each individual patient.

Since this task was performed by our IT experts Dr Aslam Nasir and Amanda Miotto, full appreciation must be given to them for their immense support and guidance for this section of the project. Without their input and knowledge, it would not have been possible for me to assess the functionality of the logistic equation that was derived from the Canterbury IBD data study. A screenshot of the clinical CD risk calculator is shown below in Figure 4.10. By simply clicking on a patients result for each predictive value using the YES/NO option, a risk percentage will be provided from this calculator. Using this risk evaluation, the selection of the most suitable form of treatment for a patient can be made.

A working model can be viewed at:

http: //genomicsrc.qern.qcif.edu.au/CD_Risk_Calculator/

by applying the user-name: **CDRisk** and using the password: **X44yL27x.**

This calculator is very basic at this current stage, as more time and effort was spent on the rest of this PhD project so that clinically useful and precise predictive factors could be derived. In future, it is hoped that by performing further similar studies on other large-scale data cohorts, the predictive probabilities of these and other factors could be identified to further improve the usefulness of this calculator.

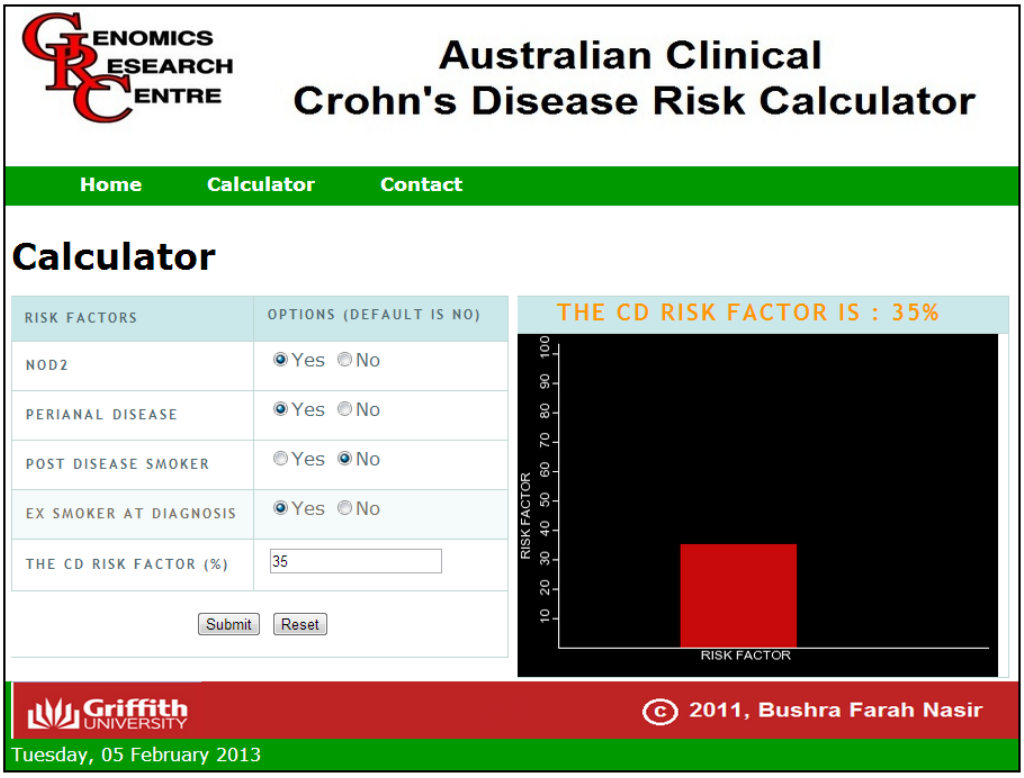

Figure 4.9 Australian Clinical CD Risk Calculator

4.6 Conclusion

The aim of this specific research was to identify personalised risk profiles to identify envirogenomic signatures that accurately predict complicated outcomes in CD patients by analysing genetic, clinical and environmental data in combination. By performing the research methods explained in this chapter, results have strongly indicated that our analysis has provided beneficial outcomes.

By performing the retrospective analysis on the Canterbury IBD cohort, we analysed a large data cohort and not only verified already identified predictive risk factors, but also showed that when these significant predictive factors are studied in combination a much higher predictive value can be obtained. The prospective analysis of the Canterbury IBD cohort further reinforced this concept, and again

highlighted that by analysing disease predictive factors in *combination* an increased and accurate prediction of the risk of surgery can be obtained.

Further to the Canterbury IBD cohort study, upon replication of these methods using the QIMR data cohort, it was once more revealed that the concept of combined factor modelling is quite useful. The CD risk calculator that has also been developed has shown that such a tool could be very useful for all types of relevant medical specialists at the time of CD diagnosis to determine an individual's risk of developing complicated, progressive forms of CD that may lead to the need for surgery, thus eliminating the possibility of unnecessary surgical interventions. Further studies, using more large-scale data cohorts, and perhaps more disease factors also, should still be performed to strengthen the methods that have been developed as part of this research study and improve the diagnostic accuracy of complicated disease prediction.

The current enthusiasm for studying genetic associations with disease, enhanced by the advent of GWAS, has tended to overshadow the important role of non-genetic (environmental) factors and the interactions that lie between genetic and non-genetic factors. While analysing comprehensive studies that involve genetic as well as non-genetic factors, standard statistical designs can be implemented to improve the power of predicting individual susceptibility towards detrimental diseases. This research study has focused on these facts and has successfully developed approaches that intend to provide personalized predictive profiles for CD patients.

CHAPTER FIVE: GENOMIC SIGNATURE DISCOVERY

***Acknowledgement*:** This chapter includes co-authored papers that have been either already published or are under review for publication, with myself as the main author. Further details regarding the papers are provided in the section titled "List of Publications and Conference Presentations". Due acknowledgment must also be made to Amanda Miotto and Dr. Aslam Nasir for their immense support with this section of the project; for their dedication to provide knowledge regarding programming issues that were faced along the way.

5.1 Introduction

Until very recently two approaches were applied to search for genes involved in the pathogenesis of CD: the 'candidate gene approach' and the 'genetic linkage study approach'. The 'candidate gene approach' investigated associations between disease and genes that are involved in the regulation of the inflammatory response. In general, differences in allelic frequencies are compared between patients and ethnically matched unrelated healthy controls. An association between the disorder and a specific marker suggests a causal relationship or, alternatively, linkage disequilibrium.

The second approach is the 'genetic linkage study' method, which does not rely on a hypothesis driven approach to identity a functional candidate gene, but uses genome wide screening in multiple affected CD pedigrees for the purpose of identifying chromosomal regions, or loci, that are shared in excess of statistical expectation in affected family members. If there is significantly increased coinheritance of polymorphic markers in the affected relatives in a given region, this region is considered to be linked to the disease. Once linkage is established the association studies of candidate genes in this region are used to identify the

specific disease gene. In recent years a new technique has evolved which has accelerated genetic research considerably. With the completion of the human genome sequence it became possible to perform GWAS, providing systematic assessment of the contribution of common variation to disease pathogenesis. This approach has had an unprecedented impact on our knowledge of the genetics of many autoimmune diseases bringing to light many unexpected candidate genes and biological pathways.

GWAS involves genotyping with several hundred thousand markers and comparing allele frequencies between patients and healthy controls. GWAS have been extremely productive, ultimately contributing to 163 independent loci for 'conventional' IBD whereas linkage studies have provided more than 13 IBD loci (Brant, 2013). The studies have resulted in the identification of many novel loci for CD and have highlighted the importance of the innate immune system and implicated new pathogenic pathways such as autophagy (Noomen *et al*, 2009). Since the development of the GWAS methods, research in the past five years has focused primarily on GWA studies which have identified susceptibility loci for over 125 complex traits in humans; however the variants discovered thus far only explain a very modest proportion of the heritability of these traits (Manolio *et al*, 2009).

Personalizing medical treatment based on the patients' genomic makeup is a major goal of clinical genetics in the twenty first century. For many common conditions a patient's health outcome (e.g. disease onset, response to treatment) is influenced by a combination of both genetic and environmental factors. Microarray technologies for high-throughput genotyping of thousands of SNPs have facilitated the GWAS approach to identifying genes related to disease

aetiology. The general logic of the GWAS is that SNP array data from case and control groups are compared and SNPs showing large allele frequency differences indicate the location of a variant gene that expresses and aberrant protein, which disrupts a biological pathway/system and results to disease. The one-SNP-at-a-time approach applied in GWAS makes it a very limited approach to identifying potential predictors of disease, and very few SNPs with ORs above 1.5 have been identified for any common human disease (Moore *et al*, 2009), thus indicating that the efficacy for GWAS for genetic testing and personalized medicine approaches will also be restricted. In spite of the ease of access to personal genetic services currently offered, the genetic architecture of common diseases is still very constrained and this knowledge has not yet guaranteed the accurate prediction for most people at risk to disease (Moore *et al*, 2009).

As a result of advances in modern technology, GWA studies have become vastly common and have helped in the identification of the differences in SNP alleles that are associated with disease. However, they have also indicated that a more multi-factor analysis approach needs to be implemented to allow a better understanding of the overall genetic makeup of diseases.

Typical GWAS techniques rely on analysing single markers individually; but it is improbable that complex diseases are caused by any single gene alone. Complex diseases are often highly heritable, but complex traits have only a small proportion of the heritability that can be explained by analysing genetic variants in traditional GWAS approaches (Manolio *et al*, 2009).

There is also the problem of 'missing-heritability', as few loci identified by GWAS have large effect sizes and it is likely that the common-disease, common-variant hypothesis (Schork *et al*, 2009, Moore *et al*, 2010) does not hold in the

case of complex diseases. GWAS also often fail to replicate the identified single marker disease associations, due to underlying epistasis (Greene *et al*, 2009). Realising the limitations of GWAS and grasping a better understanding of complex bio-molecular interactions that develop biological systems, has led to the belief that in order to map genotypes to phenotypes, a multi-gene analysis approach needs to be undertaken (Moore *et al*, 2010 and 2003).

Recently, advances in the speed, accuracy and scale of genomic sequencing improvements have also led to the first population-scale genome sequencing study (Mills *et al*, 2011). Rapid advances in DNA sequencing technology have also made WGS both technically and economically feasible (Brunham & Hayden, 2012). Full genome sequencing provides raw data on all six billion letters in an individual's DNA. However, it does not provide an analysis of what that data means or how that data can be utilized in various clinical applications, such as in medicine to help prevent disease. WGS technology is an inexpensive, time-efficient method that will eventually allow health care professionals to analyse the entire human genome of an individual and therefore detect all disease-related genetic variants, regardless of genetic variants prevalence or frequency.

The GWAS design is directed toward identifying genomic signatures of disease to overcome the limitations of the gene mapping strategy for predicting disease outcome. A genomic signature is the unique combination of genotypes from multiple unlinked SNPs that best explains (or predicts) a trait. The primary goal of genomic signature identification is to develop DNA-based tests for diagnostic use in clinical practice. This differs from the conventional gene mapping goals, which aim to hone in on a single disease susceptibility gene (or variant) and then understand its causal role in disease pathology with the view of developing better

medicines. With genomic signature identification genotypes from multiple SNPs localizing to different regions of the genome (or SNP sets) are analysed collectively for association with the trait.

By employing this multi-SNP analysis it is expected that a greater proportion of the genetic variance of disease will be explained and therefore the SNP set will be of greater predictive value. The larger effect size of a genomic signature may also mean that smaller sample sizes are required to detect disease associations in the first place. Because the genomic signature approach is aimed at disease prediction rather than causation the genomic location or the functional relevance of the SNPs is not of primary interest.

The latter does not necessary mean that genomic signatures will not be biologically important. Indeed, it is plausible that the SNPs forming genomic signatures will be amenable to a gene network analysis, which should provide valuable clues about inherited biochemical and cellular pathways that cause a disease state. By using the genomic information from GWAS and WGS methods to identify multi-SNP or genetic signatures associated with disease, it is predicted that physicians and genetic counsellors will be able to predict future disease occurrence in a person and therefore be able to minimize the impact of that disease, or may even be able to avoid the disease

To date there have been over 500 GWAS projects published for over 100 human traits and diseases and 1,698 associated SNPs (Catalogue of Published GWAS). Many GWAS data sets have been made publically available in the hope that bioinformatics scientists would be able to yield improved genetic insights which could be used to prevent and treat the disease. Scientists are now busy attempting

to determine the functional significance of the implicated SNPs and associated genes in the hope of identifying drug targets for the disease in question. The new GWAS findings have also raised hopes that genetic testing of disease-associated SNPs could be used clinically to predict an individual's genetic risk of disease and for assigning treatment accordingly (i.e. personalized medicine).

The often limited information available about environmental exposures and other non-genetic risk factors in GWA studies will make it difficult to identify gene-environment interactions or modification of gene-disease associations in the presence of environmental factors (Pearson *et al*, 2008). For personalized medicine to be acknowledged there are at least four important technical limitations of the conventional GWAS approach that need to be recognized, considered and overcome:

1. ***Heritability***

The utility of GWAS for clinical application depends on the strength of the genetic influence for the trait. Many traits have heritability (H) of less than 50%. Heritability is the proportion of the variance of a trait that exists within a population that is due to inherited genetic or DNA-based factors. Identical twins share essentially 100% of their genome and therefore twin studies are a useful method to estimate heritability of a trait in the population. High heritability of a trait usually indicates that it is influenced by genetic factors and therefore environmental factors have a comparatively modest effect for this trait e.g. height has H > 80%.

Nevertheless, for many common diseases heritability is usually less than 50%, indicating the importance of the influence of environmental factors on traits. For

those disease traits with substantially very low heritability (e.g. Parkinson's disease is estimated of having H<10%) genetic variants play a very minor role in disease susceptibility and risk assessment needs to focus on environmental factors. Therefore the effectiveness of predicting the genetic risk of patients will only be valuable for those disease traits that have the greatest heritabilities.

2. ***Single Gene Analysis***

SNPs (or genes) are usually analysed individually but any single SNP will contribute a minor effect on a trait and therefore be of low diagnostic value. The GWAS analysis that is conventionally used is limited in its identification of SNPs that represent single genes (or loci); even though it is known that for common disease traits any one disease allele will only contribute a very modest influence on the total genetic risk of this trait in the population. The total genetic risk of complex diseases is explained by multiple genetic factors and other environmental factors all acting in combination and is not limited to single risk factors. Therefore when single SNPs are considered independently they will usually provide only little information about a patient's genetic risk of a disease and therefore be of very limited clinical utility. Building genomic signatures (or profiles) based on combining risk across multiple SNPs can overcome this limitation. Similarly, by combining environmental risk factors with genetic risk factors, the prediction for the total risk of a disease can be achieved.

3. ***Study Design and Statistical Analysis***

In conventional GWAS analyses, disease-associated sampling design methods do not allow for internal validation of SNPs. This means an increase in the false discovery rate. Statistically, SNPs are identified based almost exclusively on small

P-values and/or large ORs (i.e. measure of allele frequency difference between case and control groups converted in risk). SNPs are usually only considered to be robustly associated with a disease trait if they yield a P-value of less than 10^{-5} or some per test corrected value.

Whilst using odd ratios and their P-values are reasonable for discovering SNPs associated with disease susceptibility this statistical approach is not applicable for assessing the clinical diagnostic value of SNPs. Indeed, SNPs yielding large ORs in GWAS may well be very poor in terms of clinical utility and vice-versa. More appropriate statistics are required to assess diagnostic value of disease-associated SNPs for use in clinical settings and personalized medicine. Also, another cause of false-positive associations to which GWA studies are prone is population stratification. Allele frequencies vary between population subgroups, such as those defined by ethnicity or geographic origin, and these subgroups in turn differ in their risk for disease. GWA studies may then falsely identify the subgroup-associated genes as related to disease. Genotyping error is another important cause of spurious associations that must be carefully sought and corrected (Manolio *et al*, 2009).

4. ***Clinical Communication***

GWAS findings are often difficult or not appropriate to interpret when trying to diagnose individual risk in a clinical setting. The ultimate goal of personalized medicine is to use population based GWAS results to estimate the genetic risk of individuals and communicate these findings into clinically interpretable results to make clinical decisions e.g. choice of drug therapies. As mentioned above, the strength of disease associations for SNPs identified in GWAS studies is usually

reported in terms of ORs. However, odd ratios as a measure of genetic risk are not easily interpreted by clinicians and their patients as a diagnostic test result. Moreover odd ratios are not necessarily indicative of risk especially when the disease is prevalent in the population. Ideally, a genetic test should be reported as a simple "positive" or "negative" result with an associated risk score which is interpreted as the likelihood that a patient with a positive test (i.e. possesses a high risk genomic signature) will have the disease outcome.

5.2 Specific Objectives

The specific objectives for part two of this research project are outlined below. Overall the aim was to improve existing bioinformatics methods to discover and validate a genomic signature for CD using GWAS datasets (e.g. WTCCC and dbGaP).

- Objective One: Utilize GSA method to identify genomic profiles for CD in the WTCCC data set
- Objective Two: Validate the GSA method in the CD independent dataset from dbGaP

5.3 The Wellcome Trust Case Control Consortium Study

A GWA study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease (NHGRI, 2009). Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease (NHGRI, 2009). Such studies are

particularly useful in finding genetic variations that contribute to common, complex diseases (NHGRI, 2009).

GWA studies normally require two groups of participants: people with the disease (cases) and similar people without the disease (controls) (NHGRI, 2009). After genotyping each participant, the set of markers, such as SNPs, are scanned into computers (NHGRI, 2009). Then bioinformatics is applied to survey participants' genomes for markers of genetic variation (NHGRI, 2009). If genetic variations are more frequent in people with the disease, the variations are said to be "associated" with the disease (NHGRI, 2009). The associated genetic variations are then considered as pointers to the region of the human genome where the disease-causing problem is likely to reside (NHGRI, 2009). Since the entire genome is analysed for the genetic associations of a particular disease, this technique allows the genetics of a disease to be investigated in a non-hypothesis-driven manner (NHGRI, 2009). The data cohort available from the WTCCC that was used for this section of the research project involved applying the GWAS ideologies and applying them by modifying the GSA method to suit our dataset.

A large scale GWAS of 14,000 cases and 3,000 shared controls representing seven common diseases was published in 2007, by the Wellcome Trust Case Control Consortium (WTCCC, 2007). CD was included as one of the diseases that was studied and included 906 males and 1188 females, representing a total of 2094 cases and controls. From this study a number of previously reported susceptibility loci were replicated for CD, including CARD15. Other CD-susceptibility loci identified in replication included IL23R, ATG16L1 and ZNF365. New strong association signals on chromosome 3p21, 5q33, 10q24 and

18p11 were also replicated and included SNPs around the IRGM, BSN, MST1, NKx2-3 and PTPN2 genes.

This study had some limitations which included the analysis of single SNPs that each yielded small effect sizes and which explains only a minor fraction of disease variance. This analysis also did not have a validation or testing group of individuals where the results could have been validated. Furthermore, the WTCCC study was based on SNP allele frequency analysis, which takes into consideration the population as a whole instead of genotypes which are individual units and therefore should be more useful for diagnostic application.

Our research adapts the basic GWAS design towards identifying genomic signatures of disease, with a focus on CD. This research was labelled as the Genomic Signature Analysis (GSA) study. The aim of this research was to re-analyse the publically available genotype data from the WTCCC CD case control cohorts using new advanced bioinformatics techniques designed to overcome the limitations mentioned above and to ultimately generate a multi-SNP genomic signature for predicting individual risk of CD. The idea is that the resulting genomic signature should serve future epidemiological studies of differential response to non-genetic triggers and may also help with clinical diagnosis of this disorder.

The research particularly looks at complex diseases/traits which are explained by a combination of non-genetic factors and multiple genes, where the contribution of any one gene or SNP to the disease of interest is relatively modest and only explains a small fraction of the total heritable component. Thus, the predictive value of the association between any single SNP/gene and trait will be of limited

clinical use. In order to address these limitations this research develops strategies to identify genomic signatures of disease which capture more of the genetic variation and as such better predict disease.

5.3.1 Research Methodology

This research questioned whether genetic markers when acting together could be identified from the provided genotype data and if they conferred a greater genetic influence on CD compared to previous studies. The framework of this research project was largely based on similar procedures and methods explained in the paper by the WTCCC (WTCCC, 2007), however, within this study, new bioinformatics methods were applied and only CD data was analysed.

5.3.2 Data Analysis

The data was analysed using various statistical analysis software including R, Plink and SPSS. To identify genomic signatures for CD, the genotype data from the WGA studies was analysed using bioinformatics methods that included:

1) Calculating genotype frequency differences between case and control groups among all SNPs and ranking them according to the largest difference, the smallest P-value as well model consistency;

2) And carry out an independent analysis of only single SNPs using SPSS to create a SNP panel to distinguish and identify the replicative and novel genes for a gene signature.

Publicly available WGA WTCCC data sets from the UK were transferred to a secure server at Griffith University (Southport, QLD) where further processing of

the data and multiple experimental procedures as part of the GSA method were applied.

5.3.2.1 Study Participants

A total of 2,000 cases were selected in this study for CD. Control groups included a total of 3,000 participants; 1,500 participants from the 1958 British Birth Cohort and 1,500 individuals selected from blood donors recruited for this project. CD cases were attendees at IBD clinics in and around the five centres which contributed samples to the WTCCC (Cambridge, Oxford, London, Newcastle, and Edinburgh). Ascertainment was based on a confirmed diagnosis of CD using conventional endoscopic, radiological and histopathological criteria.

All subtypes of CD were included as classified by disease extent and behaviour and the collection was not specifically enriched for family history or early age of onset. The median age of diagnosis was 26.1 years and 62% of the collection had undergone CD-related abdominal surgery. Each participating sample collection was issued unique WTCCC barcode labels and unique sample identifiers for logging information on case/control status, DNA concentration, DNA extraction method, gender, broad geographical region, and age at requirement.

5.3.2.2 Data Collection

The data has been collected by the WTCCC, and was accessed by using the Griffith University parallel computing clusters as arranged with the Genomic Research Centre's bioinformatics groups. Since data was initially collected by the WTCCC, access to the data required an application to the WTCCC for data rights of use. This required signatures from supervisors and from the Griffith University computing officers; signatures can be found on the "Data Access Agreement"

form. This entire procedure was completed by the supervisor of this project, Dr Rod Lea and approval to access the data was successfully granted.

Once the data from the WTCCC had been granted rights to use for end users at Griffith University, access was allowed to Griffith University servers to collect that data and save it. Virtual Private Network software was used to create a virtual link to the Griffith University network using the internet from any out-of-campus computer by which the servers at GU Gold Coast campus could be accessed for the data. The program 'PuTTY' was then used to display a UNIX based user interface or terminal window where using provided host name and passwords, access to the Griffith University servers to the High Performance Computers was implemented. To collect and analyse the data statistically the software programs R, Plink, MDR and SPSS were used.

5.3.2.3 Genotyping

DNA data from the individuals in this study was already collected, gathered and processed by WTCCC to produce a genome wide scan of SNPs for further use. SNP genotyping was performed with the commercial release of the GeneChip 500K arrays at Affymetrix Services Lab. A modified version of the genotyping assay developed for the 100K Mapping Array was used. Samples were then processed according to set requirements and assessed.

5.3.2.4 The Genomic Signature Analysis Method

The novel Genomic Signature Analysis (GSA) bioinformatics method that has been devised for the purposes of this study is a comprehensive, step-wise approach to analysing genomic data in order to produce a precise genomic

signature to predict the risk of disease (Figure 5.1). The initial outlay of the GSA method was developed by Lea *et al* in 2009. The original method has now been formatted to suit the objectives of this study and has been further improved with the inclusion of some modifications.

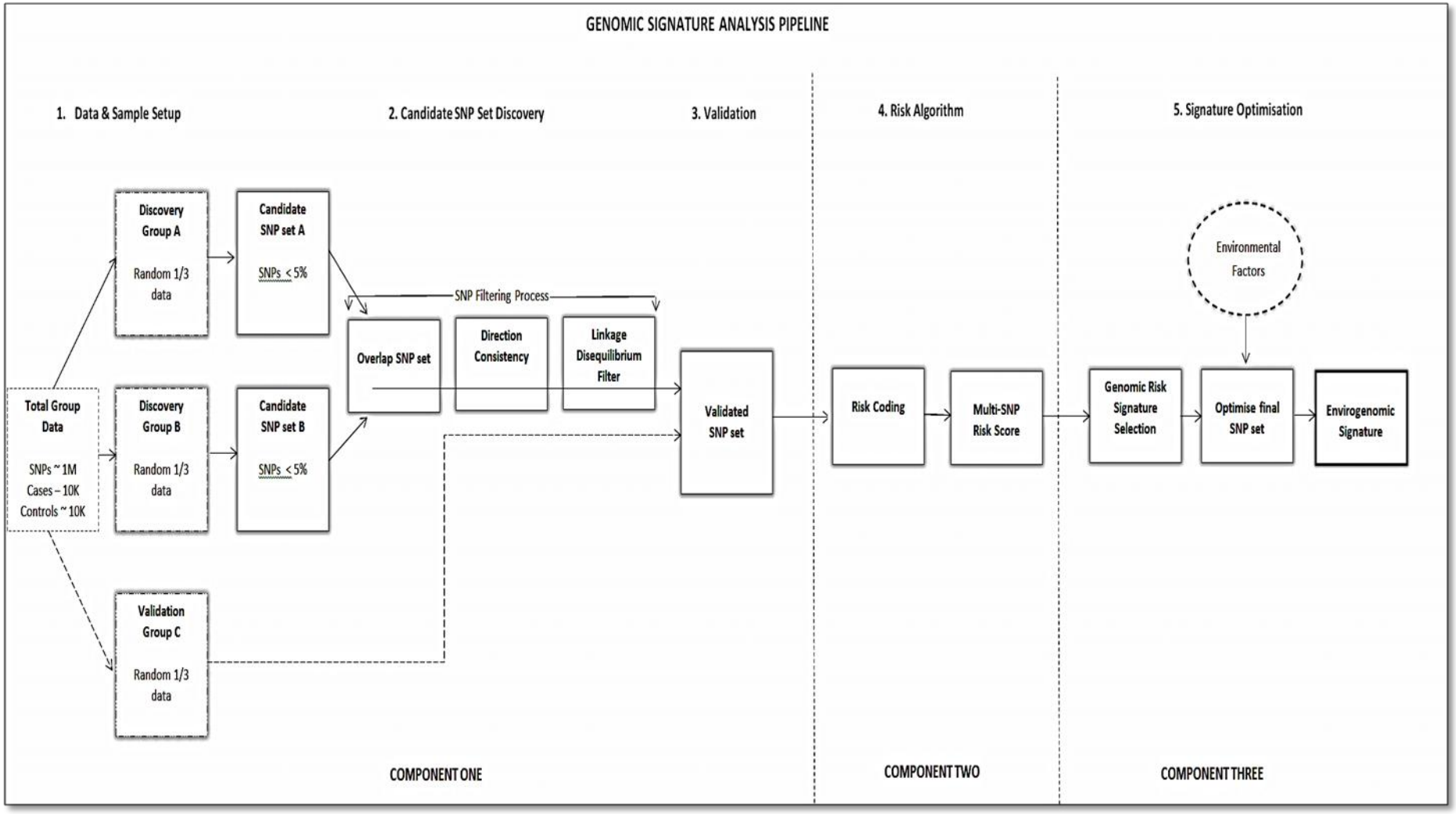

Figure 5.1 The Genomic Signature Analysis Method Flow Chart

Component One

The first component of the GSA involves a number of data cleaning and sample setup processes, a candidate SNP set discovery filtering process as well as validation of the SNP set discovered. Each step of the first component was performed 10 times to cross-validate the results obtained due to the randomisation of the data setup.

1. Data cleaning and Sample Setup

Data cleaning was performed on the entire data before it could be used for analysis so that problematic and unwanted SNPs could be excluded prior to them creating any errors in further analysis of the data and data quality control could be maintained. A discovery and validation set of the data set needed to be created from the controls and cases by randomly splitting the controls and cases. Creating these random sub-samples for both the CD files and 1958 British Birth Cohort and National Blood Service controls involved using commands in the UNIX environment on the secure servers.

The subsequently created random sub-sample files where sorted before joining. A joint controls file, which included one subsample of the 1958 British Birth Cohort file and one of the National Blood Service file, was then created. The CD and control files were then joined to create discovery and validation sets. These files were then made suitable for input into Plink. Using the created Plink input files, the data of SNP and individuals that were excluded by the WTCCC were used to create final cleaned files.

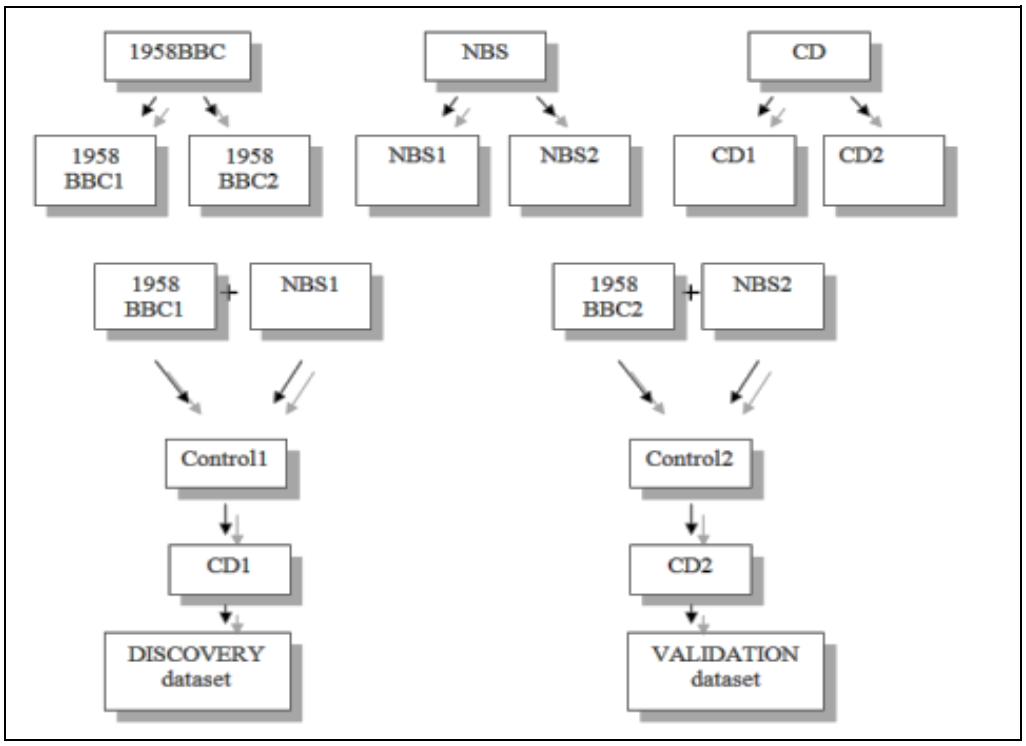

Figure 5.2 Data Cleaning Work Flow Process

2. Candidate SNP Discovery

Using both candidate SNP discovery datasets, top ranking SNPs associated with CD were selected, ranked by Chi-Square max statistics. Using R software, the maximum chi square was calculated for all SNPs. A threshold of a chi-Square maximum statistic of $\geq$ 13.8 ($P < 0.05$, 2 degrees of freedom) was selected. From these separated SNPs, consistency in both the datasets was checked and those SNPs were extracted that occurred in both the discovery and validation datasets.

After model and association tests the consistency of the models for each SNP was checked i.e. the SNPs that occurred in both model tests, and these SNPs were then extracted. With these consistent extracted SNPs, two sets of the same list of SNPs were created and on one set of SNPs linkage disequilibrium filtration was performed. The other set was left as it was and

the unfiltered and filtered files were compared for results. Genotype data from the original WTCCC file was then extracted for the SNPs from the unfiltered file. The extracted unfiltered SNPs were also compared with the WTCCC published list as well as the list for the NOD2/CARD15 gene and checked for overlaps.

3. Validation

After a set of discovery SNPs were identified, they were validated using the Validation group data file. Those SNPs that occurred in both the discovery and validation data set and were model-consistent were compared and extracted to become the final validated set of SNPs.

Component Two

The second component of the GSA involves multiplying weighted probabilities across SNPs to produce a single Genomic Risk Score (GRS). Individuals are then classified as either high risk (HR), medium risk (MR) or low risk (LR) based on their GRS. This can be done using SPSS risk score distributions by creating ROC curves and then dichotomising individuals as affected to CD (at risk) or unaffected to CD (not at risk) for each of the disease risk classifications.

Component Three

Component three of the GSA method involves the inclusion of non-genetic factors such as age and gender to further optimise the GRS. After adjusting the GRS for age and gender, final diagnostic testing can then be performed to create a final multi-factorial genomic SNP signature.

5.3.3 Results

The results obtained have been divided into the individual components of the GSA to allow for better understanding of the method itself, and the outputs that each section of the method produces.

Component One

Using the step-wise data filtration processes outlined in the first component of the GSA, a large number of SNPs are systematically analysed down to a smaller set of SNPs from each run of the GSA cross-validation process. From a total of 469,612 SNPs in every discovery file, the GSA cross-validation results are outlined in Table 5.1, explaining the number of SNPs at each step of the filtration process. The final number of validated SNPs that met the chi-square threshold criteria varies in each run due to the randomisation processes that take place during the step-wise SNP filtration processes of component one of the GSA.

Table 5.1 The Number of SNPs during each GSA Filtration Step, per Cross-Validation Run.

GSA	Discovery Overlap SNPs	Filtered, Model-Consistent Discovery SNPs	Validated Model-Consistent	Validated ChiSq> 13.8
Run 1	992	746	146	23
Run 2	748	589	98	67
Run 3	815	634	88	62
Run 4	766	596	119	79
Run 5	847	691	87	72
Run 6	893	750	140	81
Run 7	747	625	81	66
Run 8	761	607	66	55
Run 9	759	608	85	60
Run 10	837	681	100	63

The validated SNPs that emerged from each run of the GSA cross validation procedure were tested for their frequency of occurrence in each run. Those SNPs that occurred in 7 out of 10, 8 out of 10, 9 out of 10 and 10 out of 10 runs are given in Table 5.2. A total of 14 SNPs were identified that occurred in all 10 runs of the GSA after cross validation. These 14 SNPs were considered to be the most dominant, top ranking SNPs from the WTCCC data set and used for further analysis. Using the top 14 SNPs identified that occurred in 10 out of 10 runs of the GSA cross validation procedure, association tests were then performed (Table 5.3). Subsequently LD filtration was then performed, and a final set of 7 SNPs was obtained (Table 5.4). Risk statistics were then performed on this set of 7 SNPs (Table 5.5).

Table 5.2 GSA Cross Validation Top SNPs

#	7/10 runs	8/10 runs	9/10 runs	10/10 runs
1	rs10210302	rs10210302	rs10210302	rs10210302
2	rs10213846	rs10213846	rs10213846	rs11825779
3	rs10512734	rs1078621	rs11805303	rs1344485
4	rs1078621	rs11640308	rs11825779	rs1500728
5	rs11209033	rs11805303	rs11957215	rs17045918
6	rs11640308	rs11825779	rs1344485	rs2076756
7	rs11805303	rs11957215	rs17045918	rs3811417
8	rs11825779	rs12119179	rs17234657	rs3828309
9	rs11957215	rs1344485	rs204043	rs4957295
10	rs12119179	rs1505992	rs2076756	rs4957297
11	rs12325114	rs16869934	rs2201841	rs4957300
12	rs1344485	rs17045918	rs3792106	rs6431654
13	rs1503350	rs17234657	rs3811417	rs6752107
14	rs1505992	rs204043	rs3828309	rs9292777
15	rs16869934	rs2076756	rs4957295	
16	rs17045918	rs2201841	rs4957297	
17	rs17221417	rs2960920	rs4957300	
18	rs17234657	rs3792106	rs6431654	
19	rs204043	rs3811417	rs6752107	
20	rs2076756	rs3828309	rs6871834	
21	rs2201841	rs4957295	rs9292777	
22	rs2601164	rs4957297	rs943164	
23	rs2960920	rs4957300		
24	rs3792106	rs6431654		
25	rs3811417	rs6752107		
26	rs3828309	rs6871834		
27	rs4957295	rs8025932		
28	rs4957297	rs9292777		
29	rs4957300	rs943164		
30	rs6431654			
31	rs6752107			
32	rs6871834			
33	rs7170455			
34	rs8025932			
35	rs9292777			
36	rs943164			
37	rs945491			

Table 5.3 Top 14 SNP Association Test Results

*CHR: chromosome, BP: base pair, A1: first allele, F_A: Frequency affected, F_U: frequency unaffected, A2: second allele, GenoCase: Case genotypes, GenoControl: control genotypes, Trend P: Trend P value, Geno P: Genotype P Value

CHR*	SNP	BP	A1	F_A	F_U	A2	GenoCase	GenoControl	CHISQ	P	OR	Trend P	Geno P
1	rs3811417	148617980	C	0.21	0.29	T	23/615/914	41/1181/970	53.74	2.28E-13	0.67	1.88E-17	1.17E-17
2	rs10210302	233940839	C	0.401	0.48	T	279/804/615	529/1299/627	50.64	1.11E-12	0.73	5.88E-13	2.65E-13
2	rs6752107	233943448	G	0.40	0.48	A	279/803/612	529/1295/628	49.42	2.06E-12	0.73	1.13E-12	6.15E-13
2	rs6431654	233943769	C	0.40	0.48	T	277/800/611	529/1293/625	50.83	1.01E-12	0.72	5.43E-13	2.96E-13
2	rs3828309	233962410	A	0.40	0.48	G	279/809/609	529/1296/628	48.16	3.94E-12	0.73	2.08E-12	1.54E-12
4	rs17045918	114608397	C	0.22	0.15	G	62/497/845	46/518/1510	63.12	1.95E-15	1.65	1.84E-15	1.34E-14
5	rs9292777	40473705	C	0.32	0.39	T	185/710/794	386/1143/917	44.52	2.52E-11	0.73	4.89E-11	3.31E-10
5	rs4957295	40483754	A	0.26	0.33	G	121/624/938	268/1060/1110	46.63	8.57E-12	0.71	1.52E-11	9.74E-11
5	rs4957297	40490831	A	0.25	0.32	G	119/623/952	267/1057/1130	47.22	6.36E-12	0.71	1.21E-11	8.44E-11
5	rs4957300	40499496	T	0.25	0.32	C	120/623/953	268/1055/1132	46.69	8.33E-12	0.71	1.64E-11	1.14E-10
11	rs11825779	20761033	G	0.31	0.39	A	152/610/732	374/935/866	50.83	1.01E-12	0.69	6.74E-12	2.84E-11
16	rs2076756	49314382	G	0.39	0.23	A	199/665/805	74/511/842	58.74	1.80E-14	1.56	8.30E-14	1.10E-13
16	rs1344485	51469833	A	0.53	0.42	G	455/750/366	257/629/459	62.09	3.27E-15	1.52	1.59E-14	1.54E-13
23	rs1500728	28779080	T	0.52	0.47	C	615/522/561	846/632/976	14.43	0.000145	1.19	0.001571	1.56E-05

Table 5.4 Top 7 LD Filtered SNP Risk Statistics

* GenChi2: genotype Chi-Square, DomCHi2: dominant Chi-Square, RecCHi2: recessive Chi-Square, Risk GTs: risk genotypes, Common GT: common genotypes

Marker	Counts Affected		Counts Unaffected		GenChi2*		DomChi2		RecChi2		Model Consistency	Risk GTs		Common GT	
	CA_A	CA_B	CU_A	CU_B	GX2_A	GX2_B	DX2_A	DX2_B	RX2_A	RX2_B		R_High	R_Low	C_Unf	C_All
rs10210302	98/273/196	100/264/206	208/516/264	228/522/233	11.28	28.48	10.65	27.52	3.24	6.91	DOMINANT	TT	CC, CT	CT	CC
rs11825779	52/204/237	61/212/233	151/375/351	148/377/349	14.38	7.993	8.348	4.92	11.12	5.93	RECESSIVE	GA, AA	GG	GA	AA
rs1344485	158/253/115	154/251/126	87/225/146	82/199/165	21.30	26.00	12.60	20.40	15.96	14.91	MIXED	AG	AA, GG	AG	AA
rs17045918	20/173/276	21/146/297	13/216/583	19/190/639	25.98	20.16	22.57	18.88	8.40	5.30	DOMINANT	CC, CG	GG	GG	CC
rs2076756	63/227/266	69/225/268	30/169/282	28/172/271	15.33	16.53	12.04	9.96	8.19	12.08	MIXED	GA	GG, AA	AA	GG
rs3811417	7/204/302	6/203/319	19/512/374	12/494/367	40.42	44.60	40.39	44.47	0.98	0.15	DOMINANT	TT	CC, CT	CT	CC
rs4957295	44/205/313	38/209/318	94/424/464	122/422/433	10.22	25.34	10.19	20.51	1.33	12.78	DOMINANT	GG	AA, AG	GG	AA

Diagnostic testing was also performed on the final set of 7 SNPs identified (Table 5.5). All SNPs identified had high ODD RATIOS and were statistically significant.

Table 5.5 Risk Diagnostic Testing Results for the Final Top 7 SNPs

* PPV = positive predictive value, NPV = negative predictive value, OR1 = OR of cases, RR1 = relative risk of cases, RR0 = relative risk of controls

Marker	Sens	Spec	PPV	NPV	OR1	Risk	RR1	RR0
rs10210302	0.36	0.75	0.45	0.67	1.62	1.16	1.40	0.86
rs11825779	0.89	0.17	0.38	0.73	1.61	1.51	1.07	0.66
rs1344485	0.48	0.53	0.54	0.46	1.03	1.01	1.02	0.98
rs17045918	0.39	0.74	0.45	0.68	1.75	1.19	1.46	0.83
rs2076756	0.41	0.64	0.57	0.48	1.22	1.08	1.13	0.93
rs3811417	0.59	0.58	0.46	0.71	2.07	1.45	1.43	0.69
rs4957295	0.56	0.54	0.43	0.68	1.51	1.23	1.22	0.81

Component Two

Using the top 7 SNP signature identified in the first component of the GSA, a total risk score distribution was calculated and an AUC determined (Figure 5.3). A total predictive probability of 0.71 was identified *(C.I.= 0.681 – 0.734, $P=1.00x10^{-6}$, S.E. = 0.014).*

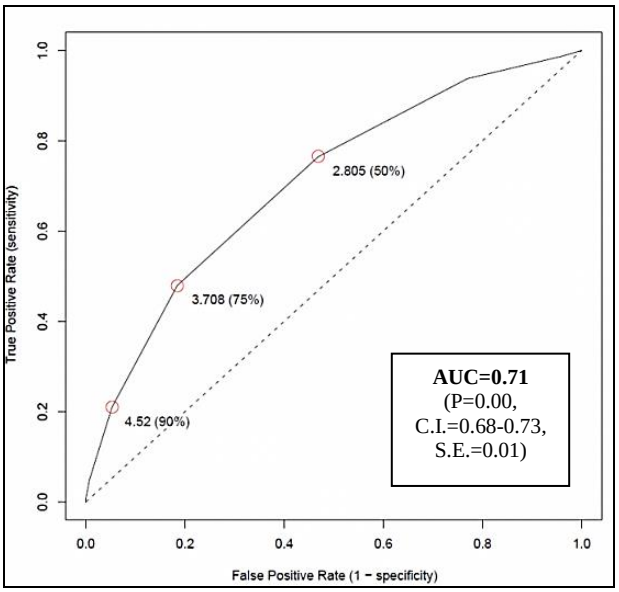

Figure 5.3 Total Risk Score Distribution of the 7-SNP Signature

The next step involved the selection of the GRS. To do this, three cut off points were selected on the ROC curve. A high risk GRS was selected at total risk ≤ 4.52, a medium risk GRS was selected at total risk ≤ 3.71 and a low risk GRS was selected at total risk ≤ 2.81 (Figure 5.4). Using the selected GRS risk distributions the dichotomisation of cases and controls is outlined in Figure 5.5.

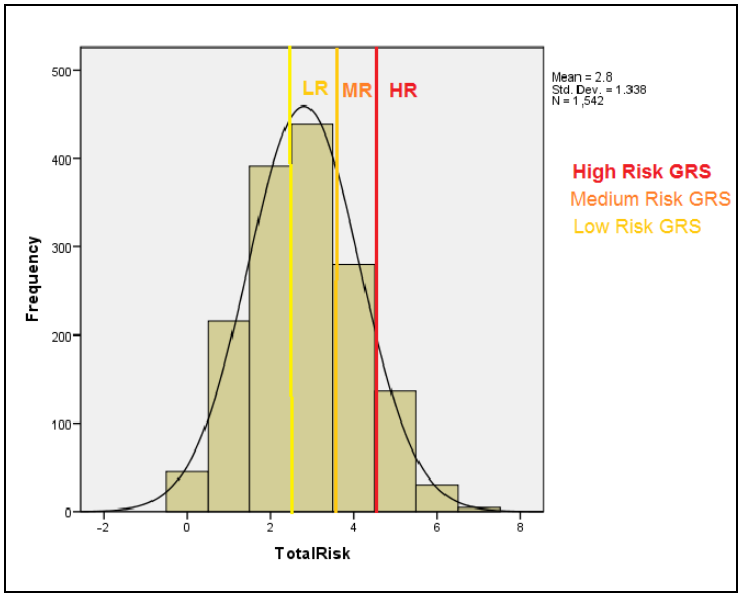

Figure 5.4 Selection of the GRS

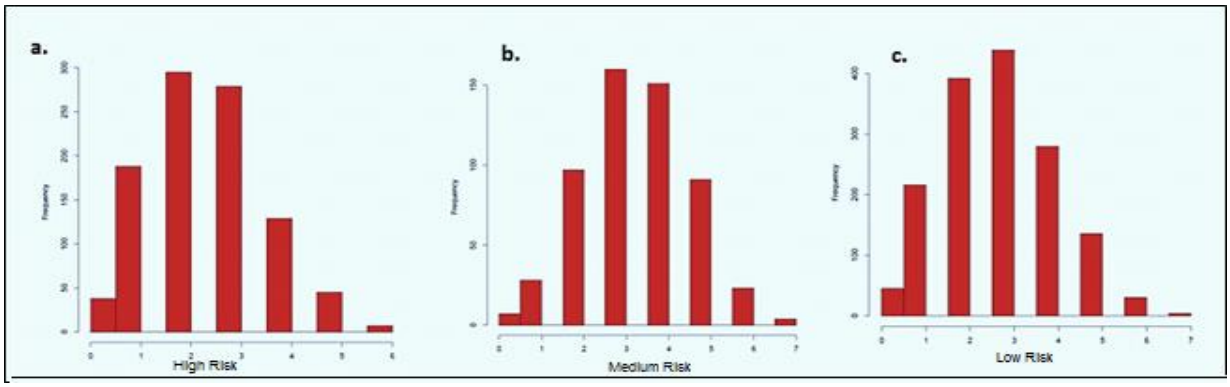

Figure 5.5 Case and Control Dichotomisations of the Selected GRS Risk Distributions:

a. High Risk GRS risk distribution of cases and controls,
b. Medium risk GRS risk distribution of cases and controls,
c. Low risk GRS risk distribution of cases and controls.

Component Three

The last component of the GSA involved diagnostic testing of the final 7 SNP GRS and also the inclusion of non-genetic factors to further enhance the 7 SNP GRS. Diagnostic testing was performed on the GRS risk distributions (Table 5.6). The high risk GRS distribution had the highest OR of 4.76 *($P = 1.0 \times 10^{-6}$)*, and the highest PPV of 0.69%. The predictive probability for the low risk GRS distribution was 0.58% *(C.I. = 0.55 – 0.61, S.E.=0.02, $P=1.0 \times 10^{-6}$)*. The medium risk GRS distribution had a predictive probability of 0.65% *(C.I. = 0.62 – 0.68, S.E. = 0.02, $P = 1.0 \times 10^{-6}$)* and the high risk GRS distribution had a predictive probability of 0.66% *(C.I. = 0.62 – 0.68, S.E. = 0.02, $P = 1.0 \times 10^{-6}$)* (Figure 5.6).

Table 5.6 Diagnostic Test Results on the GRS Risk Distributions.

GRS	OR	P Value	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Attributable Risk	Accuracy	Prevalence	Chi Square	Relative Risk
Low Risk	3.68	0.00	0.76	0.53	0.48	0.75	0.28	0.62	0.36	126.70	2.39
Medium Risk	3.94	0.00	0.48	0.81	0.59	0.73	0.32	0.96	0.36	143.41	2.21
High Risk	4.76	0.00	0.21	0.95	0.69	0.68	0.37	0.68	0.36	88.47	2.15

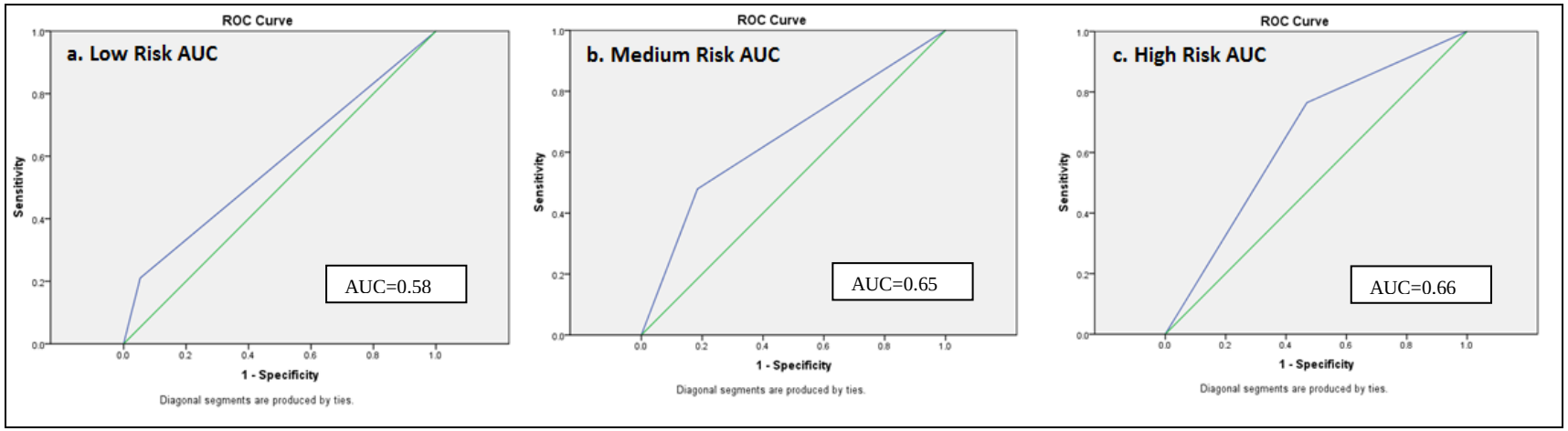

Figure 5.6 GRS AUC Distributions

Based on these results, the high risk GRS was selected as the most optimal distribution to dichotomise individuals as high or low risk, and to eventually use these results to select the most optimal form of medical treatment. The next step involved the inclusion of non-genetic factors of age and gender into the data set and then analysing the adjusted data set. When logistic regression analysis was performed on the non-genetic factors on their own both factors showed significance (Table 5.7). When these factors were tested for their diagnostic value by analysing their predictive probabilities, they both showed significance with a predictive probability of 0.76% *(C.I.=0.74–0.79, S.E.=0.02, P=1.0 x 10^{-6})* (Figure 5.7).

Table 5.7 Non-genetic Factor Unadjusted Logistic Regression Analysis

	B	**S.E.**	**P Value**	**OR**	**Lower 95% C.I.**	**Upper 95% C.I.**
Age	1.01	0.06	0.00	2.75	2.46	3.07
Gender	0.70	0.14	0.00	2.01	1.54	2.62

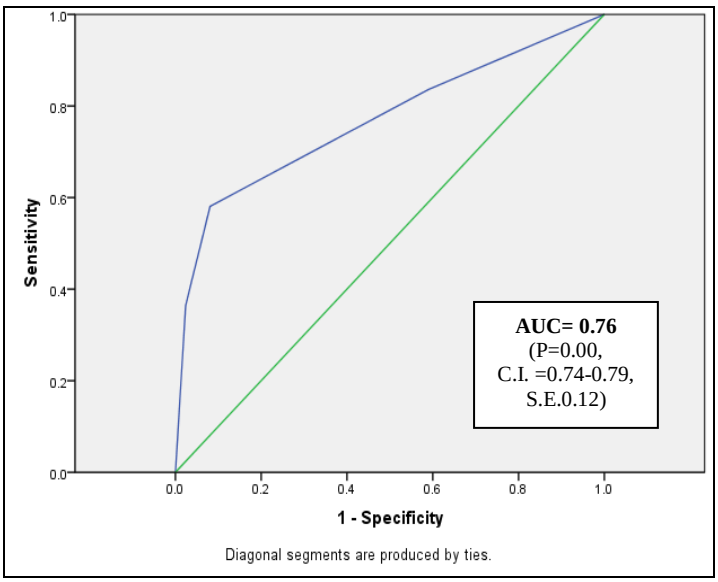

Figure 5.7: The AUC of the non-genetic factors of Age and Gender

The high risk GRS was then analysed after being adjusted for age and gender (Table 5.8). From this table we can see that all factors included were statistically significant with high ORs. Using this regression analysis, a logistic equation can also be derived (Figure 5.8).

Table 5.8 Combined Non-genetic and Genetic Logistic Regression Analysis

	B	S.E.	Sig	OR	Lower 95% C.I.	Upper 95% C.I.
Age at Recruitment	1.02	0.06	**0.00**	**2.76**	2.47	3.09
Gender	0.69	0.14	**0.00**	**2.01**	1.53	2.62
High Risk	1.02	0.19	**0.00**	**2.76**	1.91	3.99

Logit Equation = -6.024 + 1.017 x Age + .696 x Gender + 1.015 x HR distribution

Figure 5.8 Logistic Equation for the Adjusted Genomic Analysis (HR= High Risk)

To identify the predictive probability of the final factors that were analysed, a ROC curve was created (Figure 5.9). The AUC of the final predictive probability curve was 0.86 *(C.I. = 0.84 – 0.88, S.E. = 0.01, P = 1.0 x 10^{-6}).*

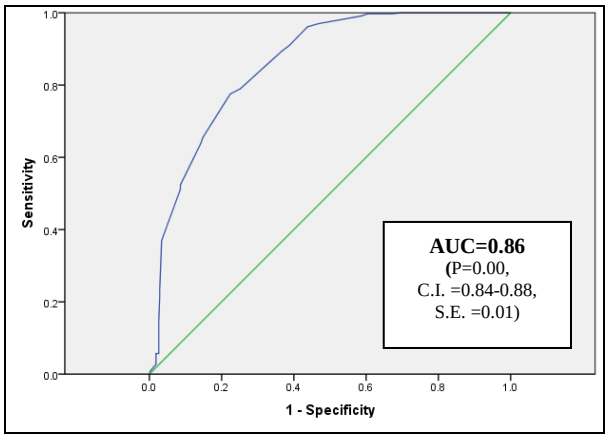

Figure 5.9 Final ROC Curve for Combined Factors.

Using these final predictive probabilities, the data can now be converted into a simple high/low test result that can indicate a positive/negative risk of CD, which can be interpreted by clinicians and patients to help diagnose and treat disease. Personalised predictive profiles can be created for patients when classified as high or low risk to CD. Using these personalised predictive profiles, clinicians can determine the best form of treatment to provide for their patients depending on their risk of CD development.

5.3.4 Discussion

Progress of the genetic research of CD has improved immensely over the past years, since the identification of the first known susceptibility gene, NOD2, in 2001. Since then, at least 71 new susceptibility genes have been identified as associated with CD, bringing the total number of genetic associations to 163. The amount of CD related genetic study research is constantly improving and given what literature is showing, it is most likely also exceeding all other research areas for CD. It can also be assumed that many more studies identifying novel susceptibility genes for CD will emerge in the coming years, as technology and laboratory techniques improve.

Besides the importance of exploring the genetic background of CD, unravelling the complicated and multifactorial pathogenesis of CD is also extremely important, if progress in developing new treatment regimens is to be made. Thus, it is the notion of this research project, that unless the disease is understood and analysed as a whole, and all influencing factors of pathogenesis are studied in *combination*, novel therapeutic treatment strategies cannot be improved to such a level that could be of the most benefit to the patient in preventing the progression of diagnosed disease to a complicated and severe condition.

Past research has shown that by combining genetic information associated with CD the risk of the development and severity of CD can be predicted (Weersma *et al.,* 2009), but studies where genetic and non-genetic or environmental data have been combined together to diagnose the risk of disease have not been previously performed. Previous studies have also shown independent predictor models for various individual risk factors (Wei *et al*, 2011) as well as formulated tools to predict the development of complicated CD using independent risk factors (Siegel *et al*, 2011). This study however, has taken a comprehensive multi-factor genomic risk model approach to identify patients at risk of CD, potentially allowing the use of early personalized therapeutic treatments to prevent the development of CD.

A number of studies have previously tackled the limitations of the single factor GWAS analysis approach to identifying genetic variants associated with disease risk using gene signature identification (Huang *et al*, 2008), multiple-loci identification (Stringer *et al*, 2011), as well as looked at multiple SNPs that lead to distinct pathways indicating disease risk (So *et al*, 2011). The potential for false-positive results, lack of information on gene function, insensitivity to rare variants and structural variants, requirement for large sample sizes, and possible biases due to case and control selection and genotyping errors, are important limitations of GWA studies (Pearson *et al*, 2008, Milne *et al*, 2008).

The often limited information available about non-genetic exposures and other non-genetic risk factors in GWA studies also make it difficult to identify gene-environment interactions or modification of gene-disease associations in the presence of non-genetic factors (Pearson *et al*, 2008). Other studies have also proven (Kang *et al*, 2011), that even when incorporating a multi-locus approach to identify the performance of disease prediction, when using only genetic data the

predictive probability of a disease can only be derived to a certain level before it begins to plateau. Since the overall risk of disease is also dependant on other non-genetic factors that work in combination with genetic factors, it makes sense to incorporate non-genetic information into the analysis when attempting to derive the best possible disease prediction models.

The current personalized genetics paradigm is developed on GWAS results but completely ignores the complexity of genotype-phenotype relationships that result from various interactions. Until the full complexity of genetic architecture is understood and accepted instead of being ignored, the era of personalized medicine will be unable to progress much further. For this to take place, researchers performing GWAS need to also incorporate gene-environment interaction data, as well as other knowledge resulting from phenomena such as epistasis, as part of their research studies.

Traditionally regression-based methods are often criticized for their inability to deal with nonlinear models and with high-dimensional data that contain many potentially interacting predictor variables, leading to sparse contingency tables that have many empty cells. For this reason, machine-learning or data-mining methods developed in the field of computer science are sometimes preferred.
The selection of predictor variables and the interactions between them that predict an outcome variable is a well-known problem in the fields of machine learning and data mining. Data-mining approaches do not fit a single pre-specified model, nor do they attempt an exhaustive search, but rather they attempt to step through the space of possible models, including potentially large numbers of main effects and multi-way interactions, in a computationally efficient way.

Many data-mining approaches are equivalent to stepping through a particular sequence of regression models and attempting to find the model that best fits the data; the distinction that is often made between data-mining and regression models is therefore, to some extent, false. One common theme in data mining is the use of cross-validation to avoid over-fitting problems (Cordell, 2009). Hence, the application of cross-validation methods incorporated in the GSA also improves its value and worth to identify possible genetic associations.

The results from this study show that the GSA method can logically and systematically extract a set of SNPs that meet a number of threshold criteria and prove to be the most significantly powerful and associated with CD. Since the GSA filtration processes are random, cross-validation of the first component of methods ensures that the resulting SNPs extracted are the most stringent and robust to be discovered.

After cross-validation of the first component, a total set of 14 SNPs were identified. After association testing and LD filtration, the set of 14 SNPs is further reduced down to a comprehensive set of 7 SNPs. After diagnostic testing on the 7 SNP set, a combined predictive probability of 0.71% *(C.I. = 0.68 – 0.73, S.E. = 0.12, P = 1.0 x 10^{-6})* was achieved. This means that those individuals with the set of 7 SNPs identified in this analysis will be at a 71% higher risk of CD than those individuals without this set of SNPs.

After selecting three risk distribution divisions for the 71% predictive probability to create a GRS, it was seen that those individuals that were dichotomised into high/low risk categories based on the High Risk GRS distribution were the most precise (PPV= 0.69), had the highest attributable risk (AR=0.37) as well as the highest specificity (0.95), along with the highest measure of effect size *(OR =*

4.76, P = 1.0 x 10^{-6}). This means that diseased individuals correctly identified as associated with CD, were the most accurately diagnosed in this GRS risk distribution. Using the high risk GRS distribution, individuals were categorised into high/low risk categories and this data used for further analysis.

By including non-genetic factors into the analysis, and adjusting the results accordingly, the predictive probability was then further enhanced. Using the high risk GRS and adjusting this data with non-genetic factors of age and gender, a final predictive probability of 0.86% *(C.I. = 0.84 – 0.88, S.E.= 0.01, P=1.0 x 10^{-6})* was achieved. This means that the final model yielded an **86%** risk of CD for those patients with the set of 7 SNPs identified from the GSA. Those patients with the 7 identified SNPs are at an 86% higher risk of developing CD than those patients without this genomic signature. This result is considerably high and can now be converted into a simple test result that can be used by clinicians to select the best method of treatment for their patients.

The GSA method outlined in this study has demonstrated a significantly important process to help classify high risk patients of CD and empower clinicians with the ability to diagnose such patients accurately and provide personalised treatment therapies accordingly. This method incorporates a step-wise systematic approach to analyse large-scale genomic cohorts to identify a single set of SNPs that portray the most association to the risk of CD. The method also highlights that with the incorporation of non-genetic factors into the analysis, the power of the predictive probability can be greatly increased. This becomes highly useful for a more accurate prediction of disease risk and allows the creation of personalised predictive profiles that can be used for a more specific, and beneficial approach to clinical treatment. From this study it can be concluded that those patients with the

7 SNP signature identified are at a much greater risk for CD, compared to those patients without this set of 7 SNPs.

Despite the plethora of genome-wide significant loci identified in CD thus far, present association signals account for only about 25% of the predicted heritability. In addition to common variation of modest effects identified through single-point analysis of the GWAS data, it is anticipated that uncommon variation at distinct loci may contribute significantly to overall disease risk. Taken together, the overall genetic architecture and optimal development of risk models of CD may be significantly more complex than previously anticipated (Kang *et al*, 2011).

A genomic signature for predicting CD will also be useful for future studies of variable drug response and gene-by-environment interactions and may eventually offer some diagnostic utility in clinical practice. This would be valuable then for the assessment of genetic and non-genetic associations that lead to the development of CD and would eventually become useful in generating DNA-based tests to diagnose this widespread, and in some cases fatal disease, and hopefully become a component of the progress for a cure in the future. This study tried to overcome some of the method limitations from the WTCCC study as well, such as applying a multi-SNP analysis and not a single SNP analysis, as well as gearing towards calculating individual risk and not just population association for CD.

Despite its appeal, genetic markers of disease will never be able to completely predict disease behaviour and severity, particularly because of the imperative role of environmental and clinical factors that influence the pathogenesis of disease. Nevertheless, genetic markers can easily be associated with other types of factors,

such as clinical or microbiological information to create more strategic disease predicting tools.

The methods that have been developed and discussed can be further enhanced by using multiple cohorts as well as including more non-genetic and clinical data in future, so that the predictive probabilities can be further improved. By applying the logic behind this method, a comprehensive tool can be created that can be used within the treatment room by clinicians to diagnose the risk of CD as soon as symptoms are detected, similar to that which has been developed in Chapter Three of this project. Patients most at risk of developing complicated disease or those requiring future surgery, can benefit from an active treatment approach using immune-suppressants or biologics, depending on an individual criteria.

5.3.5 Ethical Issues

All appropriate authorisations and clearance procedures were carried out as recommended by both Griffith University and the WTCCC. Relevant documentation was completed and submitted and clearance was granted prior to the commencement of the data analysis from the Griffith University Ethics Clearance Board. All patient participant data was kept anonymous at all times and all data obtained from the WTCCC was kept under strict surveillance and security measures.

Daily use and access of the WTCCC data from the Griffith University secure servers was done under strict controls and user access was only granted to relevant personnel. These security and safety measures were implemented throughout the entire project. This was dealt with by enforcing user names and

passwords at each step of the way to access the data from the High Performance Computer servers.

5.4 The Database of Genotypes and Phenotypes Replication Study

In order to validate the GSA method and its attempt to identify genomic signatures of disease, a replication study was carried out. To do this, data from the database of Genotypes and Phenotypes (dbGaP) was obtained. The dbGaP was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype, including GWAS. The dbGaP provides many open-access data as well as controlled-access data. The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of USA, supported by the National Institute of Health (NIH) have collected data and samples from CD patient studies as well. The study titled 'NIDDK IBDGC CD Genome Wide Association Study' was accessed and used for this research investigation.

The research objective was to evaluate the performance of the new GSA bioinformatics method on this independent GWAS dataset for CD that identifies genomic signatures for predicting disease risk. It was expected that the results of this analysis will confirm association of a genomic signature for CD. Such results should help with the diagnosis of this disorder and in so doing help personalize treatment for this debilitating disorder.

5.4.1 Research Methodology

The objective was to apply the GSA method on the independent dbGaP CD dataset to see if the discovered gene signature from the WTCCC study could be replicated. Data collection and participant recruitment procedures were all carried

out by the NIDDK. To access the CD dataset from the NIDDK, a lengthy data access application procedure was undertaken. Relevant permissions from the Griffith University signing officer, principal supervisor, IT director, and collaborators were obtained and an application was sent to the Data Access Committee. Authorized data access was approved and data was downloaded securely onto Griffith University High Performance Computer servers.

Virtual Private Network software was subsequently used to create a virtual link to the Griffith University network using the internet from any out-of-campus computer by which the servers at Griffith University Gold Coast campus could be accessed for the data. The program 'PuTTY' was then used to display a UNIX based user interface or terminal window where using provided host name and passwords, and access to the Griffith University servers to the High Performance Computers was implemented. To collect and analyse the data statistically the software programs R, Plink, MDR and SPSS were used.

5.4.2 Data Analysis

The dbGaP replication study data analysis involved a number of steps that were adapted for the GSA, and data modification procedures were carried out on the data files so that they would be appropriately executed with the GSA. A thorough study of the file structures, what type of data they contained, and how the data was formatted within the accessible files was carried out before any modifications were carried out.

5.4.2.1 Participants

Participant details have been explained from the document NIDDK data access website. Participants for the NIDDK IBDGC CD GWAS included two separate

consent groups. The first consent group were General Research Use participants of the study. They included 813 cases with both phenotype and genotype data, and 947 controls also with both phenotype and genotype data. These 1,760 individuals gave consent to general research use only for the data that was collected from them. The second consent group, labelled the 'IBD Only' group consisted of 155 cases and 48 controls with phenotype data only. These 203 individuals gave consent to use only their IBD phenotype related information.

5.4.2.2 Data Collection

The details of data collection have been extracted from the document: "*Study_Report.phs000130.IBD.v1.p1.MULTI.pdf*" provided with the data for downloading.

The dataset contained data from a GWAS performed with 968 IBD affected cases and 995 unrelated controls using the Illumina HumanHap300 Genotyping BeadChip. Cases were selected to have CD with ileal involvement, and controls were matched to cases based on gender and year of birth. Subjects were drawn from two cohorts: (*1*) persons with non-Jewish, European ancestry (561 cases and 563 controls), and (*2*) persons with Jewish ancestry (407 cases and 432 controls). Genotyping was performed at the Feinstein Institute for Medical Research.

Seven-hundred fifty-four of the samples (468 cases and 286 controls) were taken from the NIDDK IBD Genetics Consortium cell line repository and include complete phenotype data for these individuals. The remaining 1,209 samples were obtained from pre-existing collections ascertained through Cedars-Sinai Medical Centre, Johns Hopkins University, University of Chicago, University of Montreal, University of Pittsburgh, University of Toronto, and the New York Health project

(controls only). For these samples, only sex, cohort (Jewish vs. non-Jewish), and age at diagnosis (cases only) are available.

Two-hundred three individuals from among the pre-existing samples did not provide consent to release their genotype data. Thus, individual genotype data are only provided for 1,760 samples. Fifty-one samples had a call rate less than 93% and were therefore excluded from this analysis, leaving an overall sample size of 1,912 patients.

Nine samples have X chromosome heterozygosity that is neither consistent nor inconsistent with their phenotypic sex. One of these samples was found to have Turner Syndrome. The heterozygosity of the remaining 8 samples ranged from 35-76%.

5.4.2.3 Methods

The overall methods that were used to perform our replication study using dbGaP data were similar to those that were formed as part of the GSA on the WTCCC CD data set (and therefore not repeated here). Because of the differences in the formatting of files and data structure within individual files from the dbGaP, a number of data cleaning and file set-up procedures needed to be undertaken prior to implementing the GSA on the dbGaP, so that the GSA could be run effectively. There was no other major difference in the research methodology of the dbGaP replication from the initial WTCCC study.

5.4.3 Results

Due to the differences in data structure, a large proportion of time was spent in formatting the dbGaP data for use with the GSA method. At the end, the GSA

method was successfully completed on the dbGaP data set but the final WTCCC genetic signature SNPs that were previously identified were not replicated. Table 5.9 outlines the top 7 GSA SNPS identified from the WTCCC dataset and their model/association test results as identified from the dbGaP dataset.

Having found no replication of the gene signature, other SNPs that fell in close genomic regions of the top 7 SNPs were also studied to find any noteworthy findings, but it was seen that these SNPs were not of any value either (Table 5.10).

Other important factors were analysed to determine a specific cause for non-replication. Factors such as the median age at diagnosis were analysed. The dbGaP study participants had a median age at diagnosis of 23. Data available to calculate this figure was only available for 27% of participants, therefore this value was considered inaccurate to depict the median age at diagnosis for the entire dbGaP participants. As a result a statistically significant comparison between all the cohorts was not achieved. Similar outcomes were derived for other factors as well, were missing and incomplete date relayed unrealistic values for proper comparisons of all cohorts.

Table 5.9 Top 7 GSA SNP Results from dbGaP Data

CHR	SNP	BP	A1	F_A	F_U	A2	GenoCase	GenoControl	CHISQ	P	OR	Trend P	Geno P
1	rs3811417	148617980	C	0.213	0.288	T	23/615/914	41/1181/970	53.74	2.28E-13	0.6686	1.88E-17	1.17E-17
2	rs10210302	233940839	C	0.401	0.48	T	279/804/615	529/1299/627	50.64	1.11E-12	0.7253	5.88E-13	2.65E-13
4	rs17045918	114608397	C	0.221	0.147	G	62/497/845	46/518/1510	63.12	1.95E-15	1.647	1.84E-15	1.34E-14
5	rs4957295	40483754	A	0.257	0.327	G	121/624/938	268/1060/1110	46.63	8.57E-12	0.7119	1.52E-11	9.74E-11
11	rs11825779	20761033	G	0.305	0.386	A	152/610/732	374/935/866	50.83	1.01E-12	0.6984	6.74E-12	2.84E-11
16	rs2076756	49314382	G	0.318	0.230	A	199/665/805	74/511/842	58.74	1.80E-14	1.556	8.30E-14	1.10E-13
16	rs1344485	51469833	A	0.528	0.424	G	455/750/366	257/629/459	62.09	3.27E-15	1.516	1.59E-14	1.54E-13

Table 5.10 Results for 17 SNPs that Lie in Close Genomic Regions to the GSA Top 7 SNPs

CHR	SNP	BP	A1	F_A	F_U	A2	GenoCase	GenoControl	CHISQ	P	OR	Trend P	Geno P
1	rs11799813	148602364	G	0.0593	0.06091	A	3/89/709	8/99/837	0.03973	0.842	0.9719	0.8456	NA
1	rs698915	148654942	A	0.2419	0.2492	G	45/299/460	67/337/541	0.2491	0.6177	0.9614	0.6211	0.4084
2	rs6722763	233946949	G	0.1636	0.1799	A	18/229/563	33/274/638	1.627	0.2021	0.8916	0.201	0.2525
2	rs838709	233949483	C	0.2867	0.2481	A	58/349/404	60/349/536	6.64	0.009973	1.218	0.009312	0.01504
2	rs6726046	233951960	A	0.3083	0.3275	G	65/367/374	99/417/423	1.467	0.2258	0.9154	0.2176	0.2089
4	rs4834876	11468575	T	0.198	0.2083	C	32/220/465	39/289/553	0.5109	0.4747	0.9387	0.4787	0.6608
4	rs2200819	11477884	C	0.3403	0.3659	A	98/356/357	122/449/376	2.498	0.114	0.8941	0.1138	0.187
4	rs2285703	114608269	G	0.2425	0.2336	A	48/294/462	42/358/546	0.3819	0.5366	1.05	0.5308	0.3372
4	rs6835747	114610853	T	0.1547	0.155	C	23/203/579	27/239/679	0.0009	0.9761	0.9972	0.9765	0.9994
4	rs916874	114611842	C	0.2574	0.2675	A	59/300/453	67/371/506	0.4587	0.4982	0.9492	0.5012	0.5966
5	rs10512739	40467638	G	0.05916	0.06696	A	5/83/698	5/112/794	0.8651	0.3523	0.8762	0.3598	0.5248
5	rs4495224	40513272	C	0.3078	0.3494	A	88/319/397	114/433/399	6.778	0.00923	0.8283	0.01026	0.0102
11	rs8176785	20761862	G	0.2241	0.2143	A	35/293/482	39/327/579	0.489	0.4844	1.059	0.4769	0.7529
11	rs1158547	20771723	C	0.2774	0.2666	T	55/325/404	59/363/480	0.494	0.4822	1.056	0.4752	0.7722
16	rs2076756	49314382	G	0.3551	0.2444	A	127/322/362	72/318/555	51.35	7.72E-13	1.702	1.41E-11	1.09E-10
16	rs1113939	51463728	T	0.04074	0.05497	G	1/64/745	3/98/845	3.835	5.02E-02	0.7302	0.05002	NA

5.4.4 Discussion

The GSA bioinformatics method is a complex step-wise data reduction technique to identify genomic signatures of disease. It was hoped that by replicating the substantial results obtained from the GSA on the WTCCC dataset, on another independent dataset, the value of the GSA could be further improved. In this instance however, using the dbGaP dataset, the GSA method has been unable to replicate its signature findings. However, the GSA method itself has performed validating procedures to identify a robust and reliable genetic signature for CD.

The replication results may not have been possible due to any one (or more) of the reasons discussed in this chapter, and due to the limited time available for this research program, further investigations with other independent CD datasets have not yet been possible. It is hoped that in future further replicative studies with other CD datasets can be performed to truly understand the mechanism and authenticity of the GSA method.

With the advent of GWA studies, the discovery of the genetic architecture of complex human diseases has been made possible, but several limitations and weaknesses still persist. High-throughput screening technology developments along with the accessibility of database repositories of gene-disease associations have provided researchers with an abundance of data for carrying out extensive association studies.

Given the scale of data that is analysed in GWAS, it goes without saying that such studies encounter substantial statistical and computational challenges. Perhaps the most conspicuous problem lies in multiple testing concerns which arise from the numerous statistical tests performed per dataset leading to a remarkable potential

for the discovery of false-positive findings when results are not properly corrected (Greene *et al*, 2009).

In order to minimise unauthentic associations' scientists', journal editors and researchers have formulated general guidelines for performing GWAS, encouraging replication as obligatory for validation. Replication that occurs successfully for any given research study allows significant and independent verification of the results which are obtained, thus ruling out any specious associations. Unfortunately this replication requirement may filter out real associations when those associations are a part of a larger epistatic interaction or when biology is ignored (Greene *et al*, 2009).

Naturally, replication ought to be an effective gold standard for substantiating gene-disease associations because it serves as independent statistical confirmation. Unfortunately reliable replication has not been readily attainable. In a review of genetic association study literature, Hirsch horn *et al*. (2007) reveal that from 166 reported associations, only six replicated three or more times. Shriner *et al*. (2007) and Williams *et al*. (2007) also consider the success of association studies in the genome-wide era and discuss the prevalence of findings which, in this genome-wide era, fail to replicate. Many have considered why true associations may not replicate across independent data sets. The predominant explanations account for genetic heterogeneity, environmental interactions, age-dependent effects and inadequate statistical power (Lasky-Su *et al*, 2008).

One of the likely explanations by Greene *et al* (2009) is the possibility of gene-gene interactions, as a cause for this non-reproducibility. This author suggests that markers which do not successfully replicate main effects and which do not have

interaction effects in the replication sample are not necessarily without value, but they are less likely candidates for follow-up. It is possible that these markers indicate interaction with a genetic marker unmeasured in the replication sample, a situation particularly likely if few SNPs are genotyped in the replication sample. These markers can also indicate the presence of gene-environment interaction, genetic heterogeneity, spurious results, or other complex disease aetiology.

Even though replication has become a necessity for assessing statistical results from GWAS, the replication requirement may cause real genetic effects to be missed. A real result may fail to replicate statistically significant independent genetic effects in an independent dataset when allele frequencies differ and the functional polymorphism interacts with one or more other functional polymorphisms (Greene *et al*, 2009). Studies have indicated that the power to replicate the statistically significant independent main effect of one polymorphism can drop dramatically with a change of allele frequency of less than 0.1 at a second interacting polymorphism. They also showed that differences in allele frequency can result in a reversal of allelic effects where a protective allele becomes a risk factor in replication studies. Results suggest that failure to replicate an independent genetic effect may provide important clues about the complexity of the underlying genetic architecture (Greene *et al*, 2009).

Another study by Hart *et al* (2012) suggests that many candidate gene studies use intermediate phenotypes instead of disease diagnoses. They have proposed that intermediate phenotypes have simpler genetic architectures such that individual alleles account for a larger percentage of trait variance. This implies that smaller samples can be used to identify genetic associations. Their study conducted a series of 12 candidate gene analyses of acute subjective and physiological

responses to amphetamine in 99-162 healthy human volunteers and they reported an attempt to replicate these findings in over 200 additional participants' ascertained using identical methodology. They were unable to replicate any of their previous findings. The results that were obtained in this study, raise critical issues related to non-replication of candidate gene studies, such as power, sample size, multiple testing within and between studies, publication bias and the expectation that true allelic effect sizes are similar to those reported in GWAS.

Similar to the studies explained previously, and indeed many others, in order to understand why the 7-gene signature found as a result of using the GSA method on the WTCCC data set failed to replicate in the dbGaP CD dataset, a number of aspects can be highlighted.

Population structure differences:

The individuals that were recruited for the WTCCC and the dbGaP dataset both came from different population regions and there may have been a number of differences in the environmental influences that affected these individuals. It was thought that the inherent re-sampling approach would have removed any major structural effects, however it was seen that this may not have been the case. In addition, WTCCC and dbGaP are both predominantly Caucasian samples so this should not have been a major cause for difference. However, other specific environmental factors that were not collected for analysis within the databases may be present.

Phenotype differences:

CD is supposed to have an international classification criterion, and therefore it was thought that phenotypic characteristic effects would not be

present. However, it has also been seen over the extent of this research, even with an internationally used classification criteria, that patient classification of various phenotypes varies between clinicians and gastroenterologists.

Association analysis:

This was an association analysis i.e. studying correlation, not causation. Conventional GWAS are of similar design and many have successfully identified robust associations across different populations. Still, it is difficult to pass judgment as this reasoning may still be a cause for the lack of replication.

Trait has only a minor genetic component:

CD is thought to have more than just a minor heritability. However it may have been possible that the individuals that were been studied have been influenced by other non-genetic factors at a higher proportion than the genetic influence of CD.

Replication using the dbGaP may have not been possible due to any number of reasons, however to investigate the specific cause of the lack of replication would have taken quite a lot of time. The attempt at developing a significantly accurate and robust bioinformatics method to identify gene signatures associated with disease is still on-going. Even though replication has become a strong requirement in genetic association studies, so much so, that it is difficult to publish a paper without it. However, Greene's (2009) study showed that the power to replicate under an epistasis model can drop from more than 80% to less than 20% with very small changes in allele frequencies in replication data. This paper showed using

simulation, that lack of replication can provide misleading evidence in support of the null hypothesis of no association.

It can be concluded that maybe too much emphasis is being placed on statistical replication. The real value of any genetic association is whether someone is convinced to spend the money and time to experimentally validate a finding. Replication certainly helps but is not the only piece of evidence that should be considered (Moore, 2009). It is anticipated that further work and research using other independent datasets and with more time to scrutinise the analysis, a more confident outcome for the GSA method can be achieved.

5.5 Ethical Issues

Possible ethical issues that may have arisen as possible limitations or complications within this project had been vastly already covered and taken care of before the commencement of the project. Relevant authorisations and clearance procedures were requested and granted prior to the gathering of the data from participants for this project by supervisors. Regarding the rest of the project that was performed, ethical issues included accessing data from the Griffith University servers which was downloaded from dbGaP and its security and availability were all maintained throughout the project. This was dealt with by enforcing user names and passwords at each step of the way to access the data from the High Performance Computer servers.

CHAPTER SIX: CONCLUSION

6.1 Overview

Over the past many decades, the incidence of IBDs such as CD has increased worldwide and is growing continuously in almost all geographical and socioeconomic regions. Even in countries where the occurrence of IBDs was previously rare, industrialization and urbanisation are stimulating the development of IBDs worldwide. In spite of its degree of existence in the World's population though, the cause of CD still remains unidentified.

The most commonly accepted causes that influence the progression and development of CD include three main factors: genetic susceptibility, environmental influences and the clinical manifestations that transpire as a result of a homeostasis imbalance between the intestinal microbiome and host immunity. These three factors have been shown to interact with each other through complex mechanisms and are known to cause the numerous defined phenotypes of CD. Despite the progress in identifying significant phenotypes and obtaining precise insights into their characteristics, the complete pathogenesis of CD is still incomplete. The pathogenesis of CD almost certainly is mediated by a complicated network of effects, meaning that neither genetic nor environmental causes are separate reasons of the disease state, but instead interactions within and between each of these causes determine the risk of disease. Investigations aimed at clarifying these interactions ultimately will lead to a clearer understanding of the underlying progressions of CD, thus providing insight for improved disease prevention and treatment for patients.

Much of the current genetics research taking place today focuses on identifying new risk factors and pathways that lead to the development of CD. Although the potential for genomics to contribute to clinical care has long been anticipated, the pace of defining the risks and benefits of incorporating genomic findings into medical practice has been relatively slow (Manolio *et al*, 2012). Accurately diagnosing patients in time to prevent complicated and severe disease forms still remains a challenging task for clinical researchers and health care practitioners.

Consequently, it makes sense to state that further research in alleviating these challenges for the benefit of the patient needs to be undertaken. This research investigation has attempted to make use of already identified CD risk factors and create personalised predictive profiles so that patients who have a higher chance of developing severe, complicated disease forms can be treated at the time of diagnosis to control and restrict the progression of disease. The chance that an individual is at a higher risk will depend on which key CD risk factors are relevant and playing their role in the development of the disease for that individual. Identifying these risk factors for each individual has been the ultimate goal for this research project.

The ability to improve and predict outcomes for CD with individualized medical therapies generally necessitates two criteria: **1**) prediction of an individual's disease progression and **2**) means to optimize therapy. Recent evidence suggests that both criteria can be progressively accomplished, and foretelling outcomes is no longer simply relying on the gradual treatment of symptoms. Evolutions in end points for clinical trials and clinical practice goals, along with improved disease prediction and optimization of medical therapies, provide examples of how

progress in the personalization of medicine that can alter the course of these chronic immune-inflammatory disorders has taken place (Hanauer, 2012).

The sequencing of the human genome, followed by the related HapMap project, and the explosive quantity of GWA and WGA studies accomplished over the last number of years, has heralded a new era of genetics and medicine.

Personalized medicine has the potential to transform healthcare through earlier diagnosis, more effective prevention and treatment of disease, and avoidance of drug side effects. Even so, the pace of realizing the potential of genomics to contribute to clinical care has appeared slow to come, although clinical adoption of scientific discoveries has been estimated to take up to 17 years. The relatively robust genotype-phenotype associations for common complex diseases only began to become available around 2005 (Manolio *et al*, 2012) and since then, despite the immense amount of research carried out in the area of personalized medicine, there has been relatively little implementation of personalized medicine approaches into simultaneous clinical practice.

The proposed routine implementation of personalized medicine practices included the use of WGS and GWAS. At the time of the first draft of the human genome, medical revolutions of personalized and DNA-based medicine were claimed by many, and believed to be imminent. After many years now however, the much-awaited revolution has still not arrived.

The vast majority of genomic data is not medically usable. Almost all significant GWAS have identified genetic associations for common, complex diseases but confer ORs that are too small to be medically useful. These facts however do not down-play the extraordinary biological insights that have been achieved thus far.

Instead, they indicate the need for further research discoveries into investigating new avenues of personalized medicine such as how multiple risk variants combine in additive and multiplicative ways to clarify clinical diagnostic risk.

The enthusiasm that exists in current genetics research for studying genetic associations with disease tends to minimize the fundamental significance of other non-genetic factors such as environmental and clinical risks. It is understandable that investigating the interactions between genetic and non-genetic factors and describing their associations towards the risk of disease is quite complex and difficult. However, this difficult undertaking needs to be further researched if significant and essential improvements in personalized medicine are to be made.

Accordingly, the conception of this research project was to incorporate the personalised predictive profile strategy in allowing a more defined and accurate diagnosis of CD as well as identifying its future severity possibilities at the very beginning of the diagnosis and eventually incorporating this information into clinical practice. This research project has taken upon exploring CD risk variants in an effort to discover how these variants act upon disease prediction when studied in a collective mode.

The research presented in this thesis explores a number of different approaches with various formats of data to highlight and understand that the objective for this project can be implemented in multiple ways for multiple data types. For example, the first section of the project looks largely at environmental and clinical data observations, with relatively little genetic data available. The second section of the project mostly involves genetic data, and incorporates age and gender as exemplary environmental factors, into the methods investigated. The extensive

process of research represented in this chapter clearly explains the advantages of this exploration technique, especially when translating this research into a clinical setting.

The prevention of common diseases relies on identifying all possible risk factors. Instigating interventions in high-risk groups can then be possible. With advancing genetic technology, it will be possible to refine the risk-factor approach to target intervention to individuals with risk factors who also carry disease-susceptibility allele(s). Nevertheless, most known risk factors have low PPVs. The use of genetic testing can markedly increase the PPV of a risk factor and it is suggested that the use of genetic tests is likely to improve the disease-predictive value of risk factors (Khoury and Wagener, 1995).

Recent exploration into the interactions of environmental factors along with genetic factors in determining disease risks has now showed that studying gene-environment interactions may be more clinically relevant and useful for diagnostic testing. For most diseases and genetic risk factors, such cofactors are still poorly understood, and a lot of work, particularly in population-based studies, is needed before the results of basic genetic research can be translated into population-based interventions.

This research investigation also adopts the concepts behind current research methods and it took a step further by incorporating not only genetic and environmental risk factors, but clinical risk factors into the paradigm as well, so that a more well defined and accurate disease risk prediction could be achieved. Similar gene-environment interaction analyses have also been performed recently, and several methods have also been investigated. A complete study that has been

able to incorporate not only genetic factors, but environmental and clinical factors in combination, and then also produce realistically high predictive results has not been previously published.

CD is a debilitating disease that often manifests during the prime years of life. A patient's quality of life is immeasurably affected and day-to-day activities are haltered in severe complicated disease forms. In patients most at risk of having disabling disease or for those that require surgery, active treatment with immunosuppressants' or biologics should be discussed in balance with the goal of therapy, which will differ based on the patient's characteristics. For example, for an individual who is young, a non-smoker, and diagnosed with limited stricturing disease, surgical resection would be the most preferable treatment option because if such a patient is left unoperated, the disease is most likely to progress and adopt a disabling path of progression whereas surgery would be able to provide several years of remission. On the other hand, if a retired, aged patient, who smokes, has perianal disease and with severe clinical manifestations of the disease located at an ileocolonic location, the best treatment would be to prescribe immunosuppressants and biologics straight after diagnosis. Such patients would also be at a high-risk for postoperative disease recurrence and therefore surgical interventions should be avoided so as to maintain the patients' quality of life as much as possible.

High-risk patients should be regularly checked for further disease development indications and treated as quickly as possible to prevent additional clinical manifestations of disease. Prognostic tools that can predict a patient's level of risk towards development of problematic and severe disease characteristics are therefore an essential tool for medical practitioners to have at the time of disease

identification. The value of such tools would be enormous, especially from the patient's perspective, since those individuals who can avoid surgical clinical outcomes would be able to maintain a reasonably satisfactory life. This research project has intended to develop a preliminary prognostic tool that would be able to predict a patient's risk of developing severe complex CD based on the risk factors that were analysed as part of this study.

At this point it is necessary to step back and identify the key discoveries made from this research project to comprehend the value of the analytical techniques that have been undertaken as part of the project.

6.2 Envirogenomic Risk Profiling

In the envirogenomic risk profiling section of the research project two datasets were analysed to identify potential risk factors for CD that are able to predict the need for any type of IBD-related surgery. An investigation of these risk factors when they are acting in combination with each other was carried out, with the concept that they would be able to better predict the risk of requiring surgical interventions. The methods were applied on the data to be analysed in a retrospective and then also in a prospective method, and were also replicated using an independent data cohort.

6.2.1 The Canterbury IBD Retrospective Research Discoveries

The Canterbury IBD retrospective research study explored genetic and clinical data from a population based cohort from New Zealand. The data was examined in a retrospective approach by only taking those risk factors previously identified from this cohort into consideration for the analysis. The methods were able to

identify coupled associations between genetic and clinical factors that influence the risk of disease severity based on the clinically evaluated outcome of the risk for IBD-related surgery.

The study identified the NOD2 genotype (*OR=2.84, P=1.0x10*$^{-2}$) and perianal disease (*OR=2.84, P=1.0x10*$^{-2}$) as predictors of surgery in this cohort. Subsequent analysis revealed that when the genetic (NOD2) and clinical (perianal disease) factors were combined into a single risk factor and examined, the combined clinicogenetic factor had a high OR of 3.84 (*P=1.0x10*$^{-4}$) high specificity of 0.92, and reasonable PPV of 62%. Clinically, this study discovered a clinicogenetic risk factor that showed that patients who had this risk factor i.e. had the NOD2 gene and had perianal disease, were at a considerably higher risk of developing complicated CD and requiring surgical mediations.

Even when risk prediction is clinically useful, genetic risk may not be - some of those individuals with a genetic risk factor will never develop CD, while others without the genetic risk factors may develop CD due to other possible risk factors, such as clinical exposures. Therefore, this study shows that it is important to identify all (and not just genetic) predicting factors in an individual. Thus, by analysing these risk factors in a combined approach a higher overall personalized predictive probability can be obtained. This research will be more beneficial when implemented in a clinical setting for the accurate prediction of the risk of developing complicated disease for clinicians and medical practitioners, so that more individually tailored treatment regimens can be utilised for such patients with this debilitating disease.

6.2.2 The Canterbury IBD Prospective Research Discoveries

The Canterbury IBD research study also experimented with data in a prospective manner, by analysing all factors that were gathered for the Canterbury IBD project for this analysis. This research investigation undertook a systematic step-wise data reduction method to identify CD risk factors based on the need for surgery as the primary clinical outcome that was evaluated.

The research showed that after analysing genetic, clinical and environmental factors individually, a number of disease predictor factors were identified from this cohort. These factors included the NOD2 gene ($OR=1.60,\ P=1.6x10^{-3}$) perianal disease ($OR=2.695,\ P=1.0x10^{-4}$), patients who were ex-smokers at diagnosis ($OR=2.706,\ P=1.0x10^{-4}$) and post-diagnosis smokers ($OR=5.359,\ P=1.20x10^{-3}$). All these factors were significantly associated with the need for IBD-related surgery. When adjusted for age and gender, the current age of an individual also showed association to the need for IBD-related surgery ($OR=1.01,\ P=4.40 \times 10^{-3}$).

When diagnostic testing was performed on these predictive factors the ORs for all factors significantly increased. The combined multi-factor model also provided the highest predictive probability of 68% ($P=1.0 \times 10^{-4}$), greater than when any of the factors were analysed individually.

From this study it was concluded that CD patients from this New Zealand based population cohort who possess all these predicting factors are at a higher risk of requiring IBD-related surgery. From the detailed findings presented in previous chapters, it can be implied that demographically similar populations would also have similar outcomes. This research emphasises the need for a combined additive

approach of analysis when determining predictive risk factors of disease, so that clinically applicable results can be obtained and shows that the objectives for this study were successful.

6.2.3 The QIMR Replication Study Discoveries

After identifying predictive risk factors that were associated with the need for surgical interventions in CD patients, from the Canterbury IBD cohort a replicative study was performed with the QIMR IBD data cohort. The methods that were implemented in the first section of this study were repeated with this dataset and results indicated positive outcomes.

It was shown after analysis that perianal disease and smoking were both predictive factors associated with IBD-related surgery. Perianal disease had an OR of 3.14 ($P=1.0x10^{-6}$) and Smoking had an OR of 6.03 ($P=1.0x10^{-6}$). When analysed in a combined method, these predictive factors provided a substantially high predictive probability of 82% ($P=1.0x10^{-6}$), which was much higher than when any of these factors were analysed individually for their predictive probabilities.

The multi-factor predictive model identified from this study suggests that a randomly selected patient from this multi-cohort dataset would have an 82% risk of needing IBD-related surgery if they possessed both these predictor factors, compared to those patients who possess neither of these factors. Clinically, obtaining an 82% predictive probability is deemed as quite substantial, and this multi-factor model could be viable when applied in an actual clinical setting for further testing.

The multi-factor model produced offered substantial increase in the predictive value over either of the factors when considered independently and suggest that further investigative studies using these methodologies may be able to provide a fairly accurate diagnostic accuracy, which would be realistic for clinical applications.

6.2.4 Clinical Risk Calculator Development

Another objective of this part of the research project was to develop a user-friendly clinically applicable computational tool. It was endeavoured that this tool could be hypothetically utilized by medical practitioners at the time of diagnosis to predict future development of complicated CD.

Using the logistic equation derived from the Canterbury IBD research study, this calculator is able to predict the risk of disease progression based on the identified predictive factors. In future, it is hoped that more large-scale data can be analysed with vast amounts of factors, so that the calculator can be further developed by incorporating other predictive risk factors, thus providing the patient with a detailed and thorough overview of their risk of complicated disease forms.

6.3 Genomic Signature Profiling

This section of the research project involved analysing the WTCCC dataset for CD and implementing a novel step-wise data reduction algorithm at identifying genetic signatures associated with the disease. Besides the importance of exploring the genetic background of CD, unravelling the complicated and multi-factorial pathogenesis of CD is also extremely important, if progress in developing new treatment regimens is to be made. Thus, the objective of this section of the project was to understand and analyse the genetic disease

associations as a whole, and to identify a single genetic signature associated with this disease.

It was emphasised in this section of the analysis that all influencing factors of pathogenesis should be studied in combination, so that novel therapeutic treatment strategies can be improved to such a level that could be of the most benefit to the patient in preventing the progression of diagnosed disease to a complicated and severe condition. The genomic signature profiling methods that were implemented for this analysis were independently validated and provided significant outcomes. These methods were also replicated with the dbGaP data for CD, even though the GSA methods derived were already self-sufficiently verified within the method itself.

6.3.1 The WTCCC CD GSA Discoveries

The genomic signature profiling section of this research project showed that a novel GSA method is able to logically and systematically extract a set of SNPs that meet certain distinct threshold criteria and prove that the identified SNPs are the most significantly powerful in their association with CD. The method itself has been further improved from the original version derived by Lea *et al* (2009), by means of extensive investigative efforts. Various changes were gradually incorporated into the GSA throughout the analytical steps that were undertaken to improve the final outcomes. The final GSA method proved to be more accurate and clinically relevant then the original GSA methods derived by Lea *et al* in 2009.

The GSA algorithm was able to identify a 7 SNP signature for CD from the WTCCC data cohort for CD after an extensive step-wise data reduction analysis.

After diagnostic testing on the 7 SNP set, a combined predictive probability of 71% ($P=1.0x10^{-6}$) was achieved. This means that those individuals with the set of 7 SNPs identified in this analysis will be at a 71% higher risk of CD than those individuals without this set of SNPs.

After selecting three risk distribution divisions for the 71% predictive probability to create a GRS, it was seen that those individuals that were dichotomised into high/low risk categories based on the High Risk GRS distribution were the most precise with a PPV of 69%. This GRS distribution also had the highest attributable risk (*AR=0.37*) as well as the highest specificity (0.95), along with the highest measure of effect size ($OR=4.76$, $P=1.0x10^{-6}$). This means that diseased individuals fittingly recognized as suffering with CD, were the most accurately diagnosed in this GRS risk distribution.

Using the high risk GRS distribution, individuals were categorised into high/low risk categories and the data used for further analysis. By including non-genetic factors into the analysis, and adjusting the results accordingly, the predictive probability was then further enhanced. Using the high risk GRS and adjusting this data with non-genetic factors of age and gender, a final predictive probability of 86% (*C.I. =0.84-0.88*, *S.E. =0.01*, $P=1.0x10^{-6}$) was achieved. This means that the final model yielded an 86% risk of CD for those patients with the set of 7 SNPs identified from the GSA.

Those patients with the identified SNP signature are at an **86%** higher risk of developing CD than those patients without this genomic signature. This result is substantially high and can now be converted into a simple test result that can be used by clinicians to select the best method of treatment for their patients. Only

14% of individuals diagnosed with this SNP signature for CD in this cohort were inaccurately diagnosed for their risk of CD. Clinically, the application of this methodology into a medical situation could prove quite beneficial, and those patients, who have this 7 SNP signature, if identified at the time of diagnosis, could benefit from an early personalized treatment approach and avoid the risk of unnecessary surgical interventions.

6.3.2 The dbGaP Replicative Study Discoveries

The dbGaP study aimed at replicating the GSA method in an independent genetic data cohort to identify the same genetic signature previously derived from the WTCCC study. It was found that after extensive analysis and research the 7 SNP genetic signatures identified in the WTCCC study could not be replicated in the dbGaP dataset.

Nevertheless, the GSA method itself has performed independent validating procedures to identify a robust and reliable genetic signature for CD. The replication results may not have been possible due to any one (or more) of the reasons discussed in Chapter Five of this thesis. Due to the limited time available for this research program, further investigations with other independent CD datasets have not yet been possible.

This study undertook a GWAS approach to identifying a genomic signature for CD. Since its discovery, GWAS have described a few of hundred genetic variants that show statistically significant associations with a few traits. However the genes usually do not replicate in other investigations. Even when they do replicate, they never explain more than a tiny fraction of an interesting trait. In fact, classical Mendelian genetics based on family studies has identified far more

disease-risk genes with larger effects than GWAS research has so far (Miller, 2009). Many justifications can be connected to the relatively low findings from GWAS, especially when it comes to translating GWAS results into constructive medical practices. Even though the anticipation for genomics to contribute to clinical care practices is ongoing, it has been slower than initially predicted.

Despite its relative appeal, genetic markers on their own will probably never be able to fully predict CD behaviour, notably because of the major role of environmental and clinical factors in its disease pathogenesis. Although the identification of susceptibility genes is a major step provided by GWAS, many issues remain to be resolved, such as understanding signalling pathways (Noomen *et al*, 2009) epigenetic correlations, and gene-environment interactions.

The level of improvement delivered by a GWAS methodology, is another question of dumbfounding complexity, given that it would be impossible to grasp the complete genetic (or non-genetic) characteristics of diseases that are constantly evolving. Unravelling the complicated and multi-factorial pathogenesis of CD will nevertheless continue to progress as laboratory and data-mining techniques are further enhanced and the overall analysis process is quickened. Many pathways and treatments for CD would likely have not been considered, if not for the genetic discoveries reported using GWAS.

The GSA method is a step towards identifying genetic markers of disease from large-scale data cohorts, and providing a robust and clinically applicable predictive probability for disease. With the availability of data, the GSA method is able to incorporate environmental and clinical predictive risk factors as well, which further illuminates its genuine value. It is hoped that in future further replicative studies with other CD datasets can be performed to truly understand

the mechanism and authenticity of the GSA method, and that eventually clinical prognostic tools could be developed from the outcomes provided as a result of the GSA algorithm for the benefit of CD patients.

6.4 Envirogenomic Risk Predictions: What lies in the future?

Over the past few decades, the field of genetics has evolved tremendously. Many large-scale world-wide projects have enabled the accumulation of knowledge for many common, complex human diseases including CD. Likewise, the Human Genome Project has been able to expedite the identification of inherited genetic variants that represent the risk of occurrence or non-occurrence of complex human diseases.

After the completion of the International HapMap Project and with the development of technologies enabling genotyping individual DNA samples of 500,000 or more loci through GWAS, many discoveries have been made that illustrate the genetic architecture of complex diseases. More than 40 disease and human phenotype risk associations have been made through the implementation of GWAS analysis.

As a result of the advancements that have been made in the knowledge of these diseases through these discoveries, various companies have begun offering testing directly to consumers. These ‘genetic risk tests’ for specific diseases are also being marketed to physicians and not just consumers. It is logical that the availability of such tests for complex diseases would provide immense clinical, economic and social effects, given that these tests attest to be highly predictive and reasonably affordable. However, these consumers (and medical practitioners) ignore the fact that a great majority of the newly identified risk-marker alleles

confer very small relative risks and even when alleles that are associated with a modest increase in risk are combined, they usually end up with low predictive power. Initial GWAS findings for most diseases, including CD, have only explained a very small amount of the underlying genetic contribution to these diseases and much still remains to be understood about the overall pathogenesis of these diseases.

Recently, researchers have seen the implementation of genetic tests using WGS techniques that are capable of diagnosing genetic disorders in days instead of weeks. A decade ago, the accomplishment of the Human Genome Project created considerable optimism that cures for debilitating diseases such as CD would be available within no time. Still, cures generally have not emerged, however the ability to map human DNA quickly and cheaply is able to provide an upsurge of data on the genetic backgrounds of disease and this has been beneficial to a steadily growing number of individuals. Original sequencing prices stood at an amount of over $2 billion, since then the price, speed and accuracy have all tremendously improved; current technologies are able to read out individual genomes for approximately as low as $1000 per individual. Oxford Nanopore Technologies have also recently unveiled the first of a generation of tiny DNA sequencing devices that are the size of mobile phones, and many geneticists predict that they will also be useful as mobile phones. These devices may become readily available in the near future during routine doctor visits, enabling gene-therapy regimens altered to suit individuals with the added knowledge of existing environmental factors and clinical manifestations.

The clarification and interpretation of the human genome will be able to facilitate and transform medicine in the near future. The field of medicine today is quickly

accelerating from inefficient and experimental practices to data-driven and practical procedures. Very soon, treatments, diagnosis, prognosis and disease prevention will be tailored to each individual's specific genotypic and phenotypic facts.

Collectively analysing the risk variants that have been discovered have provided small increases in identifying the risk of these major diseases and their phenotypes. As more risk loci are discovered, the correlation between the actual genetic risk and predicted risk will also rise, but this important variable is only one of the factors that determine the overall risk of diseases.

Understandably, having any sort of personal medical information about the risk of disease would be beneficial to contain and treat the disease. The usefulness of genetic risk screening will definitely increase as more discoveries of risk loci are made; however until then one cannot be sure if this method of risk prediction would be actually beneficial for the patient in the long run. Given that genetic factors only contribute relatively small amounts of risk for common multi-factorial diseases such as CD, it is inherent that other non-genetic factors need to be analysed in addition to genetic factors, in order to construct a complete personalized predictive profile for any particular individual.

Knowing about the numerous implications that arise from solely relying on genetic factors to predict disease risk, it makes sense to incorporate other non-genetic predicting factors in scientific investigations. This particularly applies compellingly to CD, especially since common environmental and clinical risk factors that are sometimes linked with this disease are often not identified as

characteristic of CD and are frequently attributed to other ailments (e.g. appendectomies).

Studies involving genetic and non-genetic factors can be useful for investigations in many aspects e.g. analysing biological pathways, discovering genes that only act in particular environments or exposures that are hazardous only to genetically susceptible individuals, setting environmental safety standards, understanding heterogeneity in genetic associations across populations, predicting individual risk and changes that might result from changes in modifiable risk factors and choosing the best treatment based on a patient's genotype (Thomas, 2010).

By collectively analysing environmental, clinical and other non-genetic factors, along with individual genetic risks, the predictive probability of disease risk associations will undeniably increase. A more comprehensive and accurate risk prediction methodology can be developed by harnessing non-genetic factors into risk prediction analyses and not ignoring their contribution towards the risk of complex diseases.

Predicting individual risk of disease and potential changes of risk in relation to modifiable environmental factors is a key achievement of gene-environment interaction studies. Potentially protective or deleterious effects of environmental triggers on disease risks could depend upon genes involved in biological pathways and provide essential patient information for effective treatments. Those individuals susceptible of high disease risk as a result of their genetic background can be provided optimal treatment therapies accordingly.

Ultimately, the scope of this research is to provide understanding for future prognostic tests so that choosing the best treatment for an individual to maximise

response and/or minimize side effects based on a patient's individual predisposition can be achieved.

Personalized medicine assures to transform the practice of medicine, transform the global healthcare industry, and in due course lead to longer and healthier lives. Genomic data needs to be integrated with patient health records so that it can be interpreted by trained physicians who can put genomic insights in wider context. This goal is showing the way to redefining disease at the molecular level and integrating this data with patient clinical histories. This will allow clinical researchers to translate novel biomarkers and drug responses into improved treatments, which in turn can be tailored to each patient's genome. It is anticipated that such developments will be readily applicable to CD along with other multifaceted diseases.

6.5 Future Directions

This research project has been an immense undertaking, carried out with complete devotion and resilient passion. Each of the individual sections of this project have provided alternate pathways at utilising the simple idea behind this research, i.e. to analyse disease predicting factors in combination so that an increased predictive power can be obtained. The ultimate goal of this research is to eventually transform the results obtained into a clinical prognostic tool for diagnostic purposes, and as such a sample prototype has already been developed for future clinical trials.

The optimistic outcomes achieved as a result of this research project have provided an encouraging determination to progress with the concepts explained in this thesis. In future it is hoped that more replicative studies can be performed

with more large-scale data that incorporate a considerably higher number of predictive factors, including genetic, environmental as well as clinical data. It is also anticipated that the predictive probabilities of disease risk statistically proven in this thesis could be further improved through additional replicative analysis.

It is also anticipated that further analysis will lead to the creation of a more refined and accurate prognostic tool for diagnosis and that this tool would be able to provide beneficial insight for both medical practitioners and patients in diagnosing complicated outcomes of CD. Future objectives also include trialling this prognostic clinical tool in actual clinical environments to determine its actual accuracy and its advantages for disease prediction. The investigations explained in this thesis should assist to carry out future discoveries of disease prediction and risk factor associations more robust and clinically useful for diagnosing patients at high risk of debilitating disease. Eventually, it is hoped through further analysis, that all of the statistically thorough outcomes that are attained will be able to provide real-time results for patients when applied in actual clinical situations.

BIBLIOGRAPHY

Alvarez-Lobos M, Arostegui JI, Sans M, Tassies D, Plaza S, Delgado S, Lacy AM, Pique JM, Yagüe J, Panés J (2005). Crohn's disease Patients Carrying NOD2/CARD15 Gene Variants Have an Increased and Early Need for First Surgery due to Stricturing Disease and Higher Rate of Surgical Recurrence. *Ann Surg* 242 (5): 693-700

Anagnostides A, Hodgson HGF and Kirsner JB (1991). *Inflammatory Bowel Disease*, Chapman & Hall Medical, Melbourne

Andersson RE, Olaison G, Tysk G, Ekbom A (2003). Appendectomy is followed by increased risk of Crohn's disease, *Gastroenterol* 124(1): 40-6.

Science Daily (2012). More than 200 genes behind Crohn's disease identified, Science Daily Ltd. Retrieved from: http://www.sciencedaily.com/releases/2012/12/121213121805.htm. Retrieved on 19 January, 2012

Auffray C, Caulfield T, Khoury MK, Lupski JR, Schwab M and Veenstra T (2012). Looking back at genomic medicine in 2011.*Genome Med* 4(1): 9 DOI: 10.1186/gm308

Baert F, Caprilli R and Angelucci E (2007). Medical Therapy for Crohn's disease: top down or step up? *Dig Dis*; 25(3): 260-6.

Baines E (2004). Antibiotics could raise risk of Crohn's disease. *Gen Prac 2004*

Barrett JC, Hansoul S, Nicolae DL, *et al* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40(8): 955-62. DOI: 10.1038/ng.175. Epub 2008 Jun 29.

Baumgart D and Sandborn W (2012). Crohn's disease. *Lancet* 380(9853): 1590-605. DOI: 10.1016/S0140-6736(12)60026-9. Epub 2012 Aug 20. Review

Beaugerie L, Seksik P, Nion-Larmurier I (2006). Predictors of Crohn's disease. *Gastroenterol*; 130(3): 650-6.

Bernell O, Lapidus A and Hellers G (2000). Risk factors for surgery and postoperative recurrence in Crohn's disease. *Ann Surg* 231(1): 38-45.

Binion DG, (2010). Biologic Therapies for Crohn's disease: Update from the 2009 ACG Meeting. *Gastroenterol & Hepatol*, 61 (Suppl 1): 4-16.

Bouguen G and Peyrin-Biroulet L (2011). Surgery for adult Crohn's disease: What is the actual risk? *Gut* 60(9): 1178-81. DOI: 10.1136/gut.2010.234617. Epub 2011 May 24.

Brant SR (2013). Promises, Delivery and challenges of Inflammatory Bowel Disease risk gene discovery. *Clin Gastroenterol Hepatol*, 11(1): 22-6. DOI: 10.1016/j.cgh.2012.11.001. Epub 2012 Nov 3.

Brian F (2001). Crohn's disease: Safer and more effective way to treat Crohn's disease. Gastroenterology Week Atlanta, *Dig & Liv Dis* 33: 1

Bruining DH, Siddiki HA, Fletcher JG, Tremaine WJ, Sandborn WJ, Loftus EV Jr (2008). Prevalence of penetrating disease and extraintestinal manifestations of Crohn's disease detected with CT enterography. *Inflamm Bowel Dis* 14(12): 1701-6. DOI: 10.1002/ibd.20529.

Brunham LR and Hayden MR (2012). Whole-Genome Sequencing: The New Standard of Care? *Science*, 336(6085): 1112-3. DOI: 10.1126/science.1220967

Bulois P, Desreumaux P, Neut C, Darfeuille-Michaud A, Cortot A, Colombel JF (1999). Infectious agents and Crohn's disease. *Euro Soc Clin Micr Infect Dis*; 5(10): 601-4. DOI: 10.1111/j.1469-0691.1999.tb00415.x.

Card T, Logan RF, Rodrigues LC, Wheeler JG (2004). Antibiotic use and the development of Crohn's disease. *Gut* 53(2): 246-50.

Caserta L, Esposito I, Bossa F, Giaguinto S, Riegler G (2001). Appendectomy before Crohn's disease diagnosis is related to more surgical occurrence. *Digest Liver Dis*, 33, A130

Cavanaugh JA, Adams KR, Quak EJ, Bryce ME *et al*. (2003). CARD15/NOD2 Risk alleles in the development of Crohn's disease in the Australian population. *Ann Hum Genet* 67 (Pt 1); 35-41

Chatterjee N and Wacholder S (2008). Invited Commentary: Efficient testing of gene-environment interaction. *Am J Epidemiol* 169(2): 231–233. DOI:10.1093/aje/kwn352

Chiodini RJ (1989). Crohn's disease and the mycobacterioses: a review and comparison of two disease entities. *Clin Microbiol Rev* 2(1): 90–117.

Colombel JF, Sandborn WJ, Rutgeerts P, Enns R, Hanauer SB, Panaccione R, Schreiber S, Byczkowski D, Li J, Kent JD, Pollack PF (2007). Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the CHARM trial. *Gastroenterology* 132(1): 52-65. Epub 2006 Nov 29.

Cordell HJ (2009). Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 10(6): 392-404 DOI: 10.1038/nrg2579.

Corder, GW, Foreman, DI (2009). Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, *Wiley*, ISBN 978-0-470-45461-9

Cosnes J, Cattan S, Blain A, Beaugerie L, Carbonnel F, Parc R, Gendre JP(2002). Long-term evolution of disease behavior of Crohn's disease. *Inflamm Bowel Dis*; 8(4): 244-50.

Cosnes J, Beaugerie L, Carbonnel F, Gendre JP (2001). Smoking cessation and the course of Crohn's disease: An intervention study. *Gastroenterology* 20(5): 1093-9.

Cosnes J, Sokol H and Seksik P (2012). How to identify high-risk patients in Inflammatory Bowel Disease. *Crohn's disease and Ulcerative Colitis* 7: 713-725

Jung C, Colombel JF, Lemann M, Beaugerie L *et al* (2012). Genotype/Phenotype analyses for 53 Crohn's disease associated with genetic polymorphisms. *PloS One* 7(12): e52223. DOI: 10.1371/journal.pone.0052223. Epub 2012 Dec 27.

Cottone M, Rosseli M, Orlando A, Oliva L, Puleo A, Cappello M, Traina M, Tonelly F and Pagliaro L (1994). Smoking habits and recurrence in Crohn's disease. *Gastroenterol* 106(3): 643-8.

Crawford NPS, Colliver DW, Eichernberger MR, Funke AA, Kolodko V, Cobbs GA, Petras RE, and Galandiuk S (2007). CARD15 Genotype-phenotype relationships in a small Inflammatory Bowel Disease population with severe disease affection status. *Dig Dis Sci* 52: 2716-2724 DOI: 10.1007/s10620-006-9208-z

Danese S, Sans M, Fiocchi C (2004). Inflammatory Bowel Disease: the role of environmental factors. *Autoimmun Rev* 3(5): 394-400

Dinu I, Mahasirimongkol S, Liu Q, Yanai H, Sharaf Eldin N, Kreiter E, Wu X, Jabbari S, Tokunaga K and Yasui Y (2012). SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics. *PLoS ONE* 7(10): e43035. DOI: 10.1371/journal.pone.0043035

Dubinsky M, Kugathasan S, Kwon S, Haritunians T, Wrobel IT, Wahbeh G, Quiros A, Bahar RJ, Farrior S, Teleten N, Panikkath D, Ippoliti A (2011). A Combination of Genetic, Clinical and Immune Markers Predict the Need for Surgery in Crohn's disease. *Gastroenterol* 140 (5): S-153-S-153, 2011, DOI: 10.1016/S0016-5085(11)60621-5

Economou M and Pappas G (2008). New global map of Crohn's disease: genetic, environmental and socioeconomic correlations. *Inflamm Bowel Dis* 14(5): 709-20.

Eglinton TW and Gearry RB (2010). Clinical factors predicting disease course in Crohn's disease. *Expert Rev Clin Immunol*, 6(1): 41-5.

Eglinton T, Reilly M, Chang C, Barclay M, Frizelle F, Gearry R (2010). Ileal disease is associated with surgery for perianal disease in a population based Crohn's disease cohort. *Brit J Surg* 97(7): 1103-9. DOI: 10.1002/bjs.7031.

Ekbom A (1994). The epidemiology of Inflammatory Bowel Disease: a lot of data but little knowledge. How shall we proceed? *Inflamm Bowel Dis*; 10 Suppl 1: S32-4 Review.

Elding H, Lau W, Swallow DM, Maniatis N (2013). Refinement in Localization and Identification of Gene Regions Associated with Crohn's disease. *Am J Hum Genet*, 92(1): 107-13. DOI: 10.1016/j.ajhg.2012.11.004. Epub 2012 Dec 13

Therapeutic Guidelines Limited [electronic resource] (2007), retrieved January 14th, 2013, <http: //etg.tg.com.au.libraryproxy.griffith.edu.au/conc/tgc.htm?id=1cbdc3d425e6b266a3e96cf35ed4b589>

Eugene C (2011). The second European evidence-based Consensus on the diagnosis and management of Crohn's disease. *Clin Res Hepatol Gastroenterol* 35(8-9): 516-7. DOI: 10.1016/j.clinre.2011.06.009. Epub 2011 Aug 3.

Feagan BG, Panaccione R, Sandborn WJ, *et al* (2008). Effects of adalimumab therapy on incidence of hospitalization and surgery in Crohn's disease: results from the CHARM study. *Gastroenterology*, 135(5): 1493-9. DOI: 10.1053/j.gastro.2008.07.069. Epub 2008 Aug 3.

Feeney MA, Murphy F, Clegg AJ, Trebble TM, Sharer NM, Snook JA (2002). A case-control study of childhood environmental risk factors for the development of Inflammatory Bowel Disease. *Euro J Gastro Hepatol*, 14(5): 529-34.

Fedorak RN (2004). Is it time to re-classify Crohn's disease? *Clin Gastro*, 18 Suppl: 99-106.

Franke A, Fischer A, Nothnagel M, Becker C, *et al*. (2008). Genome-wide association analysis in sarcoidosis and Crohn's disease unravels a common susceptibility locus on 10p12.2. *Gastroenterol* 135(4): 1207-15. DOI: 10.1053/j.gastro.2008.07.017. Epub 2008 Jul 18.

Franke A, Hampe J, Rosenstiel P, Becker C, *et al*. (2007). Systematic association mapping identifies NELL1 as a novel Inflammatory Bowel Disease gene. *PLoS One* 2(8): e691

Franke A, McGovern DPB, Barret JC, Wang K*et al* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Gen*, 42(12): 1118-25. DOI: 10.1038/ng.717.

Ferguson, G A (1966). *Statistical analysis in psychology and education.* New York: McGraw-Hill

Festen EA, Goyette P, Green T, Boucher G, *et al*. (2011). A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. *PLoS Genet* 7(1): e1001283. DOI: 10.1371/journal.pgen.1001283.

Gearry RB, Lea RA, Roberts RL *et al*. (2006). CARD15 allele frequency differences in New Zealand Maori: ancestry specific susceptibility to Crohn's disease in New Zealand. *Gut* 55: 580 DOI: 10.1136/gut.2005.085464

Gearry RB, Richardson A, Frampton CM, Collett JA, Burt MJ, Chapman BA, Barclay ML. (2006). High Incidence of Crohn's disease in Canterbury, New Zealand: results of an epidemiologic study. *Inflamm Bowel Dis* 12(10): 936-43.

Gearry RB, Richardson AK, Frampton CM, Dodgshun AJ, and Barclay ML (2010). Population-based cases control study of Inflammatory Bowel Disease risk factors. *J Gastroenterol Hepatol* 25(2): 325-33. DOI: 10.1111/j.1440-1746.2009.06140.x. Epub 2010 Jan 14

Gearry RB, Roberts RL, Burt MJ, Frampton CMA, Chapman BA, Collett JA, Shirley P, Allington MDE, Kennedy MA and Barclay ML (2007). Effect of Inflammatory Bowel Disease classification changes on NOD2 genotype-phenotype associations in a population based cohort. *Inflamm Bowel Dis*, 13(10): 1220-7.

Gibson J, Collins A, Morton N (2008). Individual disease risk and multimeric analysis of Crohn's disease. *PNAS.*105; 41: 15843-15847

Greene CS, Penrod NM, Williams SM, Moore JH (2009). Failure to Replicate a Genetic Association May Provide Important Clues about Genetic Architecture. *PLoS ONE*, 4(6): e5639. DOI: 10.1371/journal.pone.0005639.

Grover S (2007). *Blueprints Pocket Gastroenterology*, Lippincott Williams and Wilkins

Gyde SN, Prior P, Macartney JC, Thompson H, Waterhouse JA, Allan RN (1980). Malignancy in Crohn's disease. *Gut* 21; 1024-9

Hahn LW, Ritchie MD and Moore JH (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19(3): 376-82.

Hanauer SB (2003). Crohn's disease: step up or top down therapy. *Best Pract Res Clin Gastroenterol*, 17(1): 131-7. Review

Hanauer SB (2012). Disease Modification in Inflammatory Bowel Disease. *Clin Gastroenterol Hepatol*, 10(9): 954-5. DOI: 10.1016/j.cgh.2012.06.011. Epub 2012 Jun 21.

Hart Al and Ng Sc (2011). Crohn's disease. *Inflamm Bowel Dis*, Medicine 39; 4: 229-236

Hart AB, Wit H, Palmer AA (2012). Candidate gene studies of a promising intermediate phenotype: Failure to Replicate. *Neuropsychopharmacol*, DOI: 10.1038/npp.2012.245. [Epub ahead of print]

Health Information Publications, (2002-2005), *Crohn's disease* [electronic resource], Health Information Publications, retrieved January 13th, 2013, Available from:

"http://www.ehealthmd.com/library/crohnsdisease/Crohn's disease_types.html"

Helbig KL, Nothnagel M, Hampe J, Balschun T, Nikolaus S, Schreiber S, Franke A and Nöthlings U (2012). A case-only study of gene-environment interaction between genetic susceptibility variants in NOD2 and cigarette smoking in Crohn's disease aetiology. *BMC Medical Genetics* 13;14

Hemminki K, Li X, Sundquist J, Sundquist K (2009). Cancer Risks in Crohn's disease patients. *Oxford J Med Ann Onc* 20(3): 574-80. DOI: 10.1093/annonc/mdn595. Epub 2008 Sep 2.

Henckaerts L, Steen KV, Verstreken I, Cleynen I, Franke A, Schreiber S, Rutgeerts P and Vermeire S (2009). Genetic Risk Profiling and Prediction of Disease Course in Crohn's disease patients. *Clin Gastroenterol Hepatol* 7(9): 972-980.e2. DOI: 10.1016/j.cgh.2009.05.001. Epub 2009 May 5.

Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA.A *Catalog of Published Genome-Wide Association Studies,* [electronic resource] Available at: www.genome.gov/gwastudies. Accessed 21st January, 2013

Hirschhorn JN (2009). Genome wide Association Studies—Illuminating Biologic Pathways. *N Engl J Med* 360(17): 1699-701. DOI: 10.1056/NEJMp0808934. Epub 2009 Apr 15.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002). A comprehensive review of genetic association studies. *Genet Med*. 4(2): 45-61.

Hodges D (2008). Top Down Approach for Crohn's safer, more effective then step-up strategy. *Medical Post Toronto*, 44; 6: 2

Hollis-Moffatt JE, Gearry RB, Barclay ML, Merriman TR, Roberts RL (2010). Consolidation of evidence for association of the KIAA1109-TENR-IL2-IL21 rs6822844 variant with Crohn's disease. *Am J Gastroenterol* 105(5): 1204-5; author reply 1206-7. DOI: 10.1038/ajg.2010.34.

Huang H, Shiffman, Friedman D, Venkatesh R, Bzowej N, Abar OT, Rowland CM, Catanese JJ, Leong DU, Sninsky JJ, Layden TJ, Wright TI, White T and Cheung R (2008). A 7 gene signature identifies the risk of developing cirrhosis in patients with chronic hepatitis C. *Hepatology* 46(2): 297-306.

Hugot JP, Alberti C, Berrebi D, Bingen E, Cézard JP (2003). Crohn's disease: the cold chain hypothesis. *The Lancet* 362(9400): 2012-5. Review

Hultén L (1988). Surgical treatment of Crohn's disease of the Small Bowel or Ileocecum. *World J Surg*; 12: 180-185

IBM (2013). SPSS Software, [electronic resource] Accessed January 17th 2013. Available from: http://www.-01.ibm.com/software/analytics/spss//

Ingle SB, Loftus EV Jr (2007). The Natural History of Perianal Crohn's disease. *Dig Liver Dis* 39(10): 963-9. Epub 2007 Aug 27.

Jantchou P, Monnet E, Carbonnel F (2006). Environmental risk factors in Crohn's disease and ulcerative colitis (excluding tobacco and appendicectomy) (Original Article in French). *Gastroenterol Clin Biol*; 30(6-7): 859-67

Johnson AD, O'Donnell CJ (2009). An Open Access Database of Genome-wide Association Results. *BMC Med. Genet*.10: 6. DOI: 10.1186/1471-2350-10-6. PMC 2639349.PMID 19161620.

Jostins L, Ripke S, Weersma RK, *et al*. (2012). Host–microbe interactions have shaped the genetic architecture of Inflammatory Bowel Disease. *Nature* 491(7422): 119-24. DOI: 10.1038/nature11582.

Kang J, Kugathasan S, Georges M, Zhao H and Cho JH (2011). Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum Mol Genet*, 20(12): 2435–2442, DOI: 10.1093/hmg/ddr116

Kappelman MD, Rifas-Shiman SL, Kleinman K, Ollendorf D, Bousvaros A, Grand RJ, Finkelstein JA (2007). The Prevalence and Geographic Distribution of Crohn's disease and Ulcerative Colitis in the United States. *Clin Gastroenterol Hepatol*, 5(12): 1424-9. Epub 2007 Sep 29.

Karlinger K, Gvörke T, Makö E, Mester A, Tarján Z, (2000). The epidemiology and pathogenesis of Inflammatory Bowel Disease. *Eur J Radiol*; 35(3): 154-67.

Khoury MJ and Wacholder S (2009). Invited Commentary: From genome-wide association studies to gene-environment interaction studies – challenges and opportunities. *Am J Epidemiol* 169(2): 227–230. DOI:10.1093/aje/kwn351

Khoury MJ and Wagener DK (1995). Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors. *Am J Hum Genet* 56(4): 835–44

Knutson D, Greenberg G, Cronau H (2003). Management of Crohn's disease—A Practical Approach. *Am Fam Physician* 68(4): 707-14.

Koltun WA (2007). *Chapter: Inflammatory Bowel Disease: Diagnosis and Evaluation*. The ASCRS textbook of colon and rectal surgery, Springer New York, pg 449

Koonin EV (2001). Computational genomics. *Curr Biol*, 11 (5): R155–8. DOI: 10.1016/S0960-9822(01)00081-1. PMID 11267880

Krishnaprasad K, Andrews JM, Lawrance IC, Florin T, *et al* (2012). Inter-observer agreement for Crohn's disease sub-phenotypes using the Montreal Classification: How good are we? A multi-centre Australasian study. *J Crohn's Colitis* 6(3): 287-93. DOI: 10.1016/j.crohns.2011.08.016. Epub 2011 Oct 4.

Kugathasan S and Amre D (2006). Inflammatory Bowel Disease- environmental modification and genetic determinants. *Pediatr Clin N Am* 53: 727-749

Kugathasan S, Maresso K, Collins N *et al* (2004). L1007Fsinc variant of CARD15/NOD2 is strongly associated with early onset and fibrostenosing behavior in pediatric Crohn's disease. *Clin Gastroenterol Hepatol* 2: 1003-9

Lacher M, Schroepf S, Ballauff A, Lohse P, Schweinitz D-V, Kappler R, Koletzko S (2009). Autophagy 16-Like 1 rs2241880 G allele is associated with Crohn's disease in German children. *Acta Peadiatrica* 98: 1835-1840, DOI: 10.1111/j.1651-2227.2009.01438.x

Laghi L, Costa S, Saibeni S, Bianchi P, Omodei P, Carrara A, Spina L, ContessiniAvesani E, Vecchi M, De Franchis R, Malesci A (2005). Carriage of CARD15 variants and smoking as risk factors for resective surgery in patients with Crohn's ileal disease. *Aliment Pharmacol Ther* 22(6): 557-64.

Lasky-Su J, Lyon HN, Emilsson V, Heid IM, Molony C, *et al* (2008). On the replication of genetic associations: Timing can be everything! *Am J Hum Genet* 82(4): 849-58. DOI: 10.1016/j.ajhg.2008.01.018.

Lakatos PL, Golovics PA, David G, Pandur T, Erdelyi Z, Horvath A, Mester G, Balogh M, Szipocs I, Molnar C, Komaromi E, Veres G, Lovasz BD, Szathmari M, Kiss LS, Lakatos L (2012). Has there been a change in the natural history of Crohn's disease? Surgical rates and medical management in a population-based inception cohort from Western Hungary between 1977-2009. *Am J Gastro* 107(4): 579-88. DOI: 10.1038/ajg.2011.448. Epub 2012 Jan 10.

Lakatos PL and Kiss LS (2010). Is the disease course predictable in Inflammatory Bowel Diseases? *World J Gastroenterol* 16(21): 2591-9.

Lakatos PL, Czegledi Z, Szamosi T, Banai J, David G, Zsigmond F, Pandur T, Erdelyi Z, Gemela O, Papp J, Lakatos L (2009). Perianal disease, small bowel disease, smoking, prior steroid or early azathioprine/biological therapy are predictors of disease behavior change in patients with Crohn's disease. *World J Gastroenterol* 15(28): 3504-10.

Legnani P and Kornbluth A (2007). Newer therapies for Inflammatory Bowel Disease. *Curr Treat Options Gastroenterol* 7(3): 161-167.

Lennard-Jones J (1989). Classification of Inflammatory Bowel Disease. *Scand J Gastroenterol* Supp 170: 2-6; discussion 16-9.

Levine JS and Burakoff R (2011). Extraintestinal manifestations of Inflammatory Bowel Disease. *Gastroenterol Hepatol* 7 (4): 235-241

Lewis CM and Knight J (2009). *Introduction to Genetic Association Studies Adapted from: Genetics of Complex Human Diseases* (ed. Al-Chalabi and Almasy) CSHL Press, Cold Spring Harbor, NY, USA

Lewis CM, Whitwell SC, Forbes A, Sanderson J, Mathew CG, Marteau TM (2007). Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn's disease. *J Med Genet*, 44(11): 689-94. Epub 2007 Jul 27.

Libioulle C, Louis E, Hansoul S, Sandor C, *et al* (2007). Novel Crohn's disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 3(4): e58. Epub 2007 Mar 5.

Lichtenstein G, Thomsen O, Schreiber S, *et al* (2009). Long-term remission with certolizumabpegol in Crohn's disease over 3.5 years: results from the PRECiSE 3 study. Presentation at the 74th American College of Gastroenterology Annual Scientific Meeting, San Diego, California 2009, Abstract 1213

Limbergen JV, Wilson DC and Satsangi J (2009). The Genetics of Crohn's disease. *Annu Rev Genomics Hum Genet* 10: 89-116

Louis E, Belaiche J, Reenaera C (2010). Do clinical factors help to predict disease course in Inflammatory Bowel Disease? *World J Gastroenterol*, 16(21): 2600-3.

Louis E, Belaiche J, Reenaera C (2009). Tailoring the treatment to the individual in Crohn's disease. *Therap Adv Gastroenterol* 2(4): 239-44. DOI: 10.1177/1756283X09337180

Louis E, Collard A, Oger AF, Degroote E, Aboul Nasr El Yafi FA, Belaiche J (2001). Behavior of Crohn's disease according to the Vienna classification: changing pattern over the course of the disease. *Gut* 49(6): 777-82.

Ludvigssona JF, Asklingb J, Ekbom A, and Montgomery SM (2006). Diagnosis underlying appendectomy and coeliac disease risk. *Ali Tract Dig Liver Dis*, 38(11): 823-8. Epub 2006 Aug 17.

Mahadevan U and Sandborn WJ (2001). Evolving medical therapies for Crohn's disease. *Curr Gastroenterol Rep*; 3(6): 471-6.

Mamula P, Markowitz JE and Baldassano RN (2003). Inflammatory Bowel Disease in early childhood and adolescence: special considerations. *Gastroenterol Clin North Am*; 32(3): 967-95, viii.

Manolio TA, Chisholm RL, Ozenberger B, Roden DM, *et al* (2013). Implementing genomic medicine in the clinic: the future is here. *Genet Med*, DOI: 10.1038/gim.2012.157 [Epub ahead of print]

Manolio TA and Collins FS (2009). The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy. *Annu Rev Med* 60: 443-56 DOI: 10.1146/annurev.med.60.061907.093117

Manolio TA, Collins FS, Cox NJ, Goldstein DB, *et al* (2009). Finding the missing heritability of complex diseases. *Nature* 61(7265): 747-53. DOI: 10.1038/nature08494.

Massouille GV, Gower-Rousseau C, Dauchet L, *et al* (2008). T1047 Gender Variations in the Incidence of Pediatric Crohn's disease: A Population-Based Cohort Study. *Gastroenterology*, 134: 4, 1; A471-A472

McGovern DP, Jones MR, Taylor KD, Marciante *et al* (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum Mol Genet*, 19(17): 3468-76. DOI: 10.1093/hmg/ddq248. Epub 2010 Jun 22.

Mills RE, Walter K, Stewart C, Handsaker RE, *et al* (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332): 59-65. DOI: 10.1038/nature09708.

Milne RL, Fagerholm R, Nevanlinna H and Benitez J (2008). The importance of replication in gene-gene interaction studies: multifactor dimensionality reduction applied to a two-stage breast cancer case-control study.

Carcinogenesis 29(6): 1215-8. DOI: 10.1093/carcin/bgn120. Epub 2008 May 14.

Molodecky NA, Soon SA, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG (2012). Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases with Time, Based on Systematic Review. *Gastroenterology*142 (1): 46 DOI: 10.1053/j.gastro.2011.10.001

Moore JH, Asselbergs F, Williams S (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26(4): 445-55. DOI: 10.1093/bioinformatics/btp713. Epub 2010 Jan 6

Moore JH (2009). From genotypes to genometypes: putting the genome back in genome wide association studies. *Eur J Human Genet* 17(10): 1205-6. DOI: 10.1038/ejhg.2009.39. Epub 2009 Mar 11.

Moore JH (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56(1-3): 73-82.

Moore JH and Williams SH (2009). Epitasis and its implications for personal genetics. *Am J Hum Genet* 85(3): 309-20. DOI: 10.1016/j.ajhg.2009.08.006.

Morris DL, Montgomery SM, Galloway ML, Pounder RE, Wakefield AJ (2001). Inflammatory Bowel Disease and laterality: is left handedness a risk? *Inflammation and Inflammatory Bowel Disease, Gut*, 49(2): 199-202.

Mosby, (1994). *Mosby's Dictionary*: Medical Nursing and Allied Health, 4th edition, Year Book Inc. Boston

Munkholm P, Langholz E, Davidsen M, Binder V (1993). Intestinal cancer risk and mortality in patients with Crohn's disease. *Gastroenterol*, 105(6): 1716-23.

Nasir B, Griffiths L, Nasir A, Roberts R, Barclay M, Gearry R and Lea R (2013). Perianal disease combined with NOD2 genotype predicts need for Inflammatory Bowel Disease-related surgery in Crohn's disease patients from a population-based cohort. *J Clin Gastroenterol* 47 (3): 242-245 DOI: 10.1097/MCG.0b013e318258314d

Neuman MG, Nanau RM (2012). Single nucleotide polymorphisms in Inflammatory Bowel Disease. *Transl Res* 160 (1): 45-64 DOI: 10.1016/j.trsl.2011.10.006. Epub 2011 Nov 23.

Nikolaus S and Schreiber S (2007). Review in Basic and Clinical Gastroenterology: Diagnostics of Inflammatory Bowel Disease. *Gastroenterol*; 133: 1670-1689

Noomen CG, Hommes DW, Fidder HH (2009). Update on genetics in inflammatory disease. *Best Pract Res Clin Gastroenterol* 23(2): 233-43. DOI: 10.1016/j.bpg.2009.02.005.

Nurmi E, Haapamäki J, Paavilainen E, Rantanen A, Hillila M and Arkkila P (2012). The burden of Inflammatory Bowel Disease on health care utilization and quality of life. *Scand J Gastroenterol* 48: 51-57 DOI: 10.3109/00365521.2012.685750

Oh S, Lee J, Kwon MS, Weir B, Ha K and Park T (2012). A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. *BMC Bioinformatics* 13 (Suppl 9): S5

Orholm M, Munkholm P, Langhold E, Neilsen OH, Sorensen TIA and Binder V (1991). Familial occurrence of Inflammatory Bowel Disease. *N Eng J Med*, 324(2): 84-8

Okazaki T, Wang MH, Rawsthorne P, Sargent M, Datta LW, Shugart YY, Bernstein CN, Brant SR (2008). Contributions of IBD5, IL23R, ATG16L1, and NOD2 to Crohn's disease risk in a population-based case-control study:

evidence of gene-gene interactions. *Inflamm Bowel Dis* 14 (11): 1528-41. DOI: 10.1002/ibd.20512

Palomino-Morales RJ, Oliver J, Gómez-García M, López-Nevot MA, Rodrigo L, Nieto A, Alizadeh BZ and Martín J (2009). Association of ATG16L1 and IRGM genes polymorphisms with Inflammatory Bowel Disease: a meta-analysis approach. *Genes Immun* 10.356-36410(4): 356-64. DOI: 10.1038/gene.2009.25

Park JH, Wacholder S, Gail MG, Peters U, Jacbos KB, Canock SJ, Chatterjee N (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet* 2(7): 570-5. DOI: 10.1038/ng.610. Epub 2010 Jun 20

Parkes M, Barrett JC, Prescott NJ, Tremelling M, *et al* (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* 39 (7): 830-2. Epub 2007 Jun 6.

Pearson TA and Manolio TA (2008). How to interpret a genome-wide association study. *JAMA*, 299(11): 1335-44. DOI: 10.1001/jama.299.11.1335.

Quezada SM, Steinberger EK and Cross RK (2012). Association of age at diagnosis and Crohn's disease phenotype. *Age and Ageing* 42: 102-1-6 DOI: 10.1093/ageing/afs107

Ramadas AV, Gunesh S, Thomas GA, Williams GT, Hawthorne AB (2010). Natural History of Crohn's disease in a population-based cohort from Cardiff (1986-2003): a study of changes in medical treatment and surgical resection rates. *Gut*, 59(9): 1200-6. Epub 2010 Jul 21.

Pugazhendhi S, Sahu MK, Subramanian V, Pulimood A, Ramakrishna BS (2011). Environmental factors associated with Crohn's disease in India. *Indian J Gastroenterol* 30(6): 264-9. DOI: 10.1007/s12664-011-0145-1. Epub 2011 Dec 13.

Ramanan VK, Shen L, Moore JH and Saykin AJ (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in Genet* 28: 7; 323-332. DOI: 10.1016/j/tig/2012/03/004

Ringel Y and Drossman DA (2001). Psychosocial aspects of Crohn's disease. *Surg Clin North Am* 81(1): 231-52

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, *et al.* (2007). Genome-wide association study identifies new susceptibility loci for Crohn's disease and implicates autophagy in disease pathogenesis. *Nat Genet* 39(5): 596-604. Epub 2007 Apr 15.

Roberts RL, Gearry RB, Hollis-Moffatt JE, Miller AL, Reid J, Abkevich V, Timms KM, Gutin A, Lanchbury JS, Merriman TR, Barclay ML and Kennedy MA (2007). IL23R R381Q and ATG16L1 T300A are strongly associated with Crohn's disease in a study of New Zealand Caucasians with Inflammatory Bowel Disease. *Am J Gastroenterol* 102(12): 2754-61. Epub 2007 Sep 25.

Roberts RL, Hollis-Moffatt JE, Gearry RB, Kennedy MA, Barclay ML and Merriman TR (2008). Confirmation of association of IRGM and NCF4 with ileal Crohn's disease in a population based cohort. *Genes Immun* 9(6): 561-5. DOI: 10.1038/gene.2008.49. Epub 2008 Jun 26.

Roberts RL, Topless RKG, Phillips-Green AJ, Gearry RB, Barclay ML and Merriman TR(2010). Evidence of interaction of CARD8 rs2043211 with NALP3 rs35829419 in Crohn's disease. *Genes Immun* 11(4): 351-6. DOI: 10.1038/gene.2010.11. Epub 2010 Feb 25.

Russell RK, Drummond HE, Nimmo EE, Anderson N, Smith L, Wilson DC, Gillett PM, McGrogan P, Hassan K, Weaver LT, Bisset M, Mahdi G, Satsangi J (2005). Genotype–phenotype analysis in childhood-onset Crohn's

disease: NOD2/CARD15 variants consistently predict phenotypic characteristics of severe disease. *Inflamm Bowel Dis* 11(11): 955-64.

Rutgeerts P (2009). Adalimumab induces and maintains mucosal healing in patients with moderate to severe ileocolonic Crohn's disease: first results of the EXTEND trial. Presentation at the 2009 Digestive Disease Week: Abstract 751e

Sandborn WJ, Gasink C, Gao LL *et al*(2012). Ustekinumab induction and maintenance therapy in refractory Crohn's disease. *N Engl J Med*367(16): 1519-28. DOI: 10.1056/NEJMoa1203572.

Satsangi J, Silverberg MS, Vermeire S, and Colombel JF (2006). The Montreal Classification of Inflammatory Bowel Disease: controversies, consensus and implication. *Gut* 55(6): 749-53.

Scherl EJ, Dubinsky M (2010). *The Changing World of Inflammatory Bowel Disease: Impact of generation, gender and global trends*. Slack Books Inc.

Schreiber S, Khaliq-Kareemi M, Lawrance IC, *et al.* (2007). Maintenance therapy with certolizumabpegol for Crohn's disease. *N Engl J Med*. 357(3): 239-50. Erratum in: N Engl J Med. 2007 Sep 27; 357(13): 1357.

Schork N, Murray S, Frazer K, Topol E (2009). Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19(3): 212-9. DOI: 10.1016/j.gde.2009.04.010. [Epub 2009 May 28].

Seiderer J, Brand S, Herrmann KA, Schnitzler F, Hatz R, Crispin A, Pfennig S, Schoenberg SO, Göke B, Lohse P, Ochsenkuhn T (2006). Predictive value of the CARD15 variant 1007fs for the diagnosis of intestinal stenosis and the need for surgery in Crohn's disease in clinical practice: results of a prospective study. *Inflamm Bowel Dis* 12(12): 1114-21.

Selby WS (2003). Current Issues in Crohn's disease. *Med J Aust* 178(11): 532-3.

Shanahan F (2002). Crohn's disease. *Lancet* 359(9300): 62-9. Review

Shriner D, Vaughan LK, Padilla MA, and Tiwari HK (2007). Problems with Genome-Wide association studies. *Science* 316(5833): 1840-2.

Siegel CA, Fleshner P, Siegel LS, Rotter Ji *et al* (2011). Predicting Crohn's disease post-operative recurrence using clinical, endoscopic, serologic and genetic factors. *Gastroenterology* 140; 5: S-153-S-153, DOI: :10.1016/S0016-5085(11)60622-7

Siegel CA, Siegel LS, Hyams JS, Kugathasan S*et al* (2011). Real-time tool to display the predicted disease course and treatment response for children with Crohn's disease. *Inflamm Bowel Dis*, 17(1): 30-8. DOI: 10.1002/ibd.21386. Epub 2010 Sep 1.

Singhal R, Taylor J, Owoniyi M, El-Khayat RH, Tyagi SK and Corfield AP (2010). The role of appendicectomy in the subsequent development of Inflammatory Bowel Disease: a UK-based study. *Int J Colorectal Dis* 25(4): 509-13. DOI: 10.1007/s00384-009-0865-1. Epub 2009 Dec 15.

So H, Gui AH, Cherny S and Sham P (2011). Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet Epidemiol* 35(5): 310-7. DOI: 10.1002/gepi.20579. Epub 2011 Mar 3.

Somerville KW, Logan RF, Edmond M, Langman MJ (1984). Smoking and Crohn's disease. *Br Med J* (Clin Res Ed) ;289(6450): 954-6.

Stenson WF and Korzenik J (2003). *Textbook of Gastroenterology; Chapter 83 Inflammatory Bowel Disease*, Lippincott Williams and Wilkins

Stokkers PCF and Hommes DW (2006). Novel Biological Therapies for Inflammatory Bowel Disease. *Curr Treat Options Gastroenterol* 9(3): 201-10.

Stone MA, Mayberry JF, Baker R (2003). Prevalence and management of Inflammatory Bowel Disease: a cross-section study from central England. *Eur J Gastroenterol Hepatol* 15: 1275-80

Stringer S, Wray NR, Kahn RS, and Derks EM (2011). Underestimated effect sizes in GWAS: Fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS ONE* 6(11): e27964. DOI: 10.1371/journal.pone.0027964. Epub 2011 Nov 28.

Takahashi H, Ando T, Watanabe O, *et al*. (2007). Usefulness of an elemental diet in Crohn's disease. *Inflammopharmacology*15 (1): 15-17

Tarrant KM, Barclay ML, Frampton CM, Gearry RB (2008). Perianal disease predicts changes in Crohn's disease phenotype – results of a population-based study of Inflammatory Bowel Disease phenotype. *Am J Gastroenterol* 103(12): 3082-93. DOI: 10.1111/j.1572-0241.2008.02212.x.

The International HapMap Consortium, The International HapMap Project, Accessed on 14th January 2013, Available from: "http://www.hapmap.ncbi.nlm.nih.gov"

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145): 661-78.

Thirlby RC, Marco A, Sobrino, MD, James B. Randall, RN (2001). The Long-term Benefit of Surgery on Health-Related Quality of Life in Patients with Inflammatory Bowel Disease. *Arch Surg* 136(5): 521-7.

Thomas D (2010). Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet* 11(4): 259-272 DOI: 10.1038/nrg2764

Thompson NP, Driscoll R, Pounder RE, and Wakefield AJ (1996). Genetics versus environment in Inflammatory Bowel Disease: results of a British twin study. *BMJ* 312(7023): 95-6.

Török HP, Glas J, Lohse P, and Folwaczny C (2006). Genetic variants and the risk of Crohn's disease: what does it mean for future disease management? *Expert Opin Pharmacother* 7(12): 1591-602.

Travis SPL, Stange EF, Lemann M *et al* (2006). European evidence based consensus on the diagnosis and management of Crohn's disease: current management. *Gut* 55: i16-i35 DOI: 10.1136/gut.2005.081950b

Tysk C, Lindberg E, Jamerot G, Floderus-Myrhed B (1988). Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*; 29(7): 990-996.

U.S. Department of Energy Genome Program's Biological and Environmental Research Information Systems (BERIS) (2011). [electronic resource] *The Human Genome Project Information*, Accessed on 15th January, 2013. Available from: "http://www.ornl.gov/hgmis"

Van Kruiningen HJ (1995). On the use of antibiotics in Crohn's disease. *J Clin Gastroenterol* 20(4): 310-6. Review

Venter J, Adams, MD, Myers, EW, Li, PW, Mural, RJ, Sutton, GG, Smith, HO, Yandell, M *et al*. (2001). The sequence of the human genome. *Science* 291 (5507): 1304–51. Bibcode 2001Sci...291.1304V. DOI: 10.1126/science.1058040.PMID 11181995.

Vermeire S, Assche VG and Rutgeerts P (2006). Review article: altering the natural history of Crohn's disease – evidence for and against current therapies. *Aliment Pharm Therap*25 (1): 3-12.

Visscher PM, Brown MA, McCarthy MI, Yang J (2012). Five years of GWAS discovery, *Am J Hum Genet* 90(1): 7-24. DOI: 10.1016/j.ajhg.2011.11.029

Walsh A, Mabee J and Trivedi K (2011). Inflammatory Bowel Disease. *Prim Care* 38(3): 415-32; vii. DOI: 10.1016/j.pop.2011.06.001. Review

Wang, McLeod LHL, and Weinshilboum R (2011). Genomics and Drug Response. *N Eng J Med* 364(12): 1144-53. DOI: 10.1056/NEJMra1010600.

Weersma RK, Stokkers PC, van Bodegraven AA, van Hogezand RA, *et al*(2009). Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. *Gut* 58(3): 388-95. DOI: 10.1136/gut.2007.144865. Epub 2008 Sep 29.

Wei SC, Ni YH, Yang HI, Su YN, Chang MC, Chang YT, Shieh MJ, Wang CY, Wong JM. (2011). A hospital-based study of clinical and genetic features of Crohn's disease. *J Formos Med Assoc*110(9): 600-6. DOI: 10.1016/j.jfma.2011.07.009. Epub 2011 Aug 19.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145): 661-78.

Williams H, Walker D, Orchard TR (2008). Extra-intestinal manifestations of Inflammatory Bowel Diseases. *Curr Gastroenterol Rep* 10(6): 597-605.

Williams SM, Canter JA, Crawford DC, Moore JH, Ritchie MD (2007). Problems with Genome-Wide association studies, *Science* 316(5833): 1840-2.

Wilson J, Hair C, Knight R, Catto-Smith A, Bell S, Kamm M, Desmond P, McNeil J Connel W (2010). High incidence of Inflammatory Bowel Disease in Australia: A prospective population-based Australian incidence study. *Inflamm Bowel Dis* 16(9): 1550-6. DOI: 10.1002/ibd.21209.

Woodcock J (2010). Assessing the clinical utility of diagnostics used in drug therapy. *Clin Pharmacol Therap* 88(6): 765-73. DOI: 10.1038/clpt.2010.230. Epub 2010 Oct 27.

Yamazaki K, McGovern D, Ragoussis J, Paoliccu M *et al* (2005). Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* 14 (22): 3499-3506 DOI: 0.1093/hmg/ddi379

Yang CH, Ding J, Gao Y, Chen X, Yang ZB, Xiao SD(2011). Risk factors that predict the requirement of aggressive therapy among Chinese patients with Crohn's disease. *J Dig Dis* 12(2): 99-104. DOI: 10.1111/j.1751-2980.2011.00484.x.

Zhang H-F, Qiu L-X, Chen Y, Zhu W-L, Mao C, Zhu L-G, Zheng M-H, Wang Y, Lei L, Shi J (2009). ATG16L1 T300A polymorphism and Crohn's disease susceptibility: evidence from 13, 022 cases and 17, 532 controls. *Hum Genet* 125: 627–631DOI 10.1007/s00439-009-0660-7

Zoccali M and Fichera A (2012). Minimally invasive approaches for the treatment of Inflammatory Bowel Disease. *World J Gastroenterol* 18 (46): 6756-6763 DOI: 10.3748/wjg.v18.i46.6756

APPENDIX A

A manuscript titled "*Perianal disease combined with NOD2 genotype predicted need for Inflammatory Bowel Disease-related surgery in Crohn's disease patients from a population-based cohort*" was published in the Journal of Clinical Gastroenterology (Nasir *et al*, 2013) on the 1st of March, 2013. This journal article was first-authored by myself and was a result of the work that is presented as part of this. A print version of this journal article is provided for perusal herewith.

Paper nor published here in order to comply with copyright