BLOOD-BASED BIOMARKERS OF ASTHMA

by

Chen Xi Yang

B.Sc., The University of British Columbia, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Experimental Medicine)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2017

© Chen Xi Yang, 2017
Abstract

Asthma is a chronic inflammatory airway disorder characterized by reversible airway obstruction and hyperresponsiveness. It affects more than 300 million people worldwide but remains poorly understood. Asthma can be induced by a variety of triggers, including allergens, substances found in the workplace, cold air and exercises, and greatly diminishes the life quality of patients. In this thesis, I investigated two major categories of asthma: allergic asthma and occupational asthma.

Two types of response are involved in allergic asthma. While 50% of the affected individuals only develop an acute early asthmatic response (EAR), the other individuals develop both the EAR and a chronic late asthmatic response (LAR). Those individuals who present an isolated EAR are classified as early responders (ERs) and individuals who present both the EAR and the LAR are classified as dual responders (DRs). In our study, patients with mild asthma were challenged with specific allergens and their blood was collected prior to the challenge. By measuring the gene expression and the protein levels of complement and coagulation molecules, I demonstrated that the complement and coagulation system may play a role in the LAR of allergic asthma.

Occupational asthma is caused by sensitivity to specific molecules found in the working environment. Western red cedar asthma (WRCA) is the most common form of occupational asthma in the Pacific Northwest region of North America, including British Columbia. It is due to sensitivity to a low molecular weight molecule, plicatic acid (PA), found in the dust of western red cedar wood. The current diagnosis of WRCA is through multiple bronchial challenges, which are time-consuming, complicated and expensive. Blood samples were collected from individuals
who were suspected to have WRCA prior to the bronchial challenges. Gene expression was measured using the NanoString platform. Using a pathway-directed approach of random forest and leave-one-out cross-validation, I identified and validated a blood-based two-gene biomarker panel which may help distinguish patients with WRCA from those with asthma due to non-western red cedar causes at baseline. Having such a biomarker panel may greatly simplify the current diagnosis of WRCA by way of a simple blood test.
Lay Summary

Asthma is a chronic inflammatory airway disorder characterized by reversible airway obstruction and affects more than 300 million people worldwide. In this thesis, I investigated two major categories of asthma: allergic asthma and occupational asthma. Allergic asthma is induced by a variety of allergens. Using a human allergen inhalation challenge model on individuals with mild asthma, I demonstrated that the complement and coagulation systems may play a role in the late asthmatic response. Similar to allergic asthma, occupational asthma is caused by specific substances found in the workspace. Using blood samples collected from individuals suspected of having western red cedar asthma and statistical modelling, I identified blood-based biomarkers which may help to diagnose this occupational disease. My findings support an approach, whereby a blood-based biomarker panel could be used to simplify the diagnosis of western red cedar asthma, and allow patients avoid the much more invasive bronchial challenges.
Preface

I participated in the experimental design and performed all the experiments and analyses described in this thesis. To distinguish from the main text, code pieces and file names in Chapter 4 will appear in a different font.

Chapters 3 and 4 contain material that has either been published in a peer-reviewed journal or a book.

Chapter 3 contains the following manuscript:


I designed the experiment and selected the study participants from discovery to validation. I processed the collected blood samples and performed RNA extraction and RNA quality assessment. I performed the NanoString assay with YWK’s help. I performed the statistical and computational analyses with AS’s help and wrote the entire manuscript. CC participated in patient recruitment and sample collection. AS, EMC, CC and SJT participated in the research design of the study.

Chapter 4 contains parts of the following book chapter:


I developed the reported RNA-sequencing data processing pipeline and wrote part of the book chapter.
Table of Contents

Abstract .......................................................................................................................... ii
Lay Summary ................................................................................................................. iv
Preface .......................................................................................................................... v
Table of Contents ......................................................................................................... vi
List of Tables ................................................................................................................ xi
List of Figures ............................................................................................................... xii
List of Symbols and Abbreviations .............................................................................. xiv
Acknowledgements ...................................................................................................... xvi
Dedication ...................................................................................................................... xviii

Chapter 1: Introduction ................................................................................................. 1

1.1 Thesis overview ........................................................................................................ 1

1.2 Allergic asthma ........................................................................................................ 2

1.2.1 Introduction ........................................................................................................ 2

1.2.2 Clinical diagnosis ............................................................................................. 2

1.2.3 Biphasic responses ......................................................................................... 3

1.2.4 Disease mechanisms ......................................................................................... 4

1.3 Occupational asthma .............................................................................................. 5

1.3.1 Western red cedar asthma .............................................................................. 5

1.3.2 Clinical diagnosis ............................................................................................. 6

1.3.3 Disease mechanisms ......................................................................................... 6

1.4 High throughput molecular technologies ............................................................... 8
1.4.1 RNA-Seq.................................................................................................................. 8
1.4.2 NanoString technologies............................................................................................ 8
1.5 Thesis Summary.............................................................................................................. 9

Chapter 2: Allergic asthma and the complement system.................................................. 11
2.1 Introduction.................................................................................................................... 11
2.2 Hypothesis...................................................................................................................... 13
2.3 Methods........................................................................................................................ 14
2.3.1 Human studies........................................................................................................... 14
2.3.2 Inhalational challenges............................................................................................... 14
2.3.3 Blood collection.......................................................................................................... 15
2.3.4 Study participants....................................................................................................... 15
2.3.4.1 Discovery set........................................................................................................... 15
2.3.4.2 Validation set.......................................................................................................... 16
2.3.5 ELISA set.................................................................................................................... 17
2.3.6 Experimental techniques............................................................................................ 18
2.3.6.1 RNA extraction ..................................................................................................... 18
2.3.6.2 RNA-Seq................................................................................................................. 19
2.3.6.3 NanoString nCounter Elements assay.................................................................. 19
2.3.6.4 ELISA ..................................................................................................................... 20
2.3.7 Data analysis............................................................................................................... 20
2.3.7.1 RNA-Seq processing............................................................................................... 20
2.3.7.2 NanoString data processing ................................................................................. 21
2.3.7.3 Differential expression analysis ........................................................................... 22
Chapter 3: Biomarkers of western red cedar asthma ........................................32

3.1 Introduction ........................................................................................................ 32
3.2 Hypothesis / Rationale ....................................................................................... 32
3.3 Methods ................................................................................................................ 33
   3.3.1 Inhalational challenges .................................................................................. 33
   3.3.2 Blood collection ............................................................................................ 34
   3.3.3 Study participants ......................................................................................... 34
      3.3.3.1 Discovery set ......................................................................................... 34
      3.3.3.2 Validation Set ......................................................................................... 35
   3.3.4 Experimental techniques .............................................................................. 36
      3.3.4.1 Blood RNA preparation .......................................................................... 36
      3.3.4.2 NanoString nCounter Elements assay ................................................. 36
   3.3.5 Statistical methodologies ............................................................................ 37
      3.3.5.1 Data normalization and batch correction ............................................. 37
      3.3.5.2 Comparisons on CBC/diffs .................................................................. 37
4.2.4.3 Running the pipeline ................................................................. 63
4.2.5 Structure of working directory ....................................................... 65
4.3 Conclusions ....................................................................................... 66
Reference ............................................................................................... 67
Appendices ............................................................................................. 72

Appendix A Supplementary material for Chapter 2 .................................. 72
A.1 Genes of interest (microarrays) comparing ER and DR at pre-challenge (8ERs, 6DRs) ................................................................. 72
A.2 Proteins of interest (iTRAQ proteomics) comparing ER and DR at pre-challenge (4ERs, 4DRs) ................................................................. 73
A.3 NanoString quality control (QC) criteria ............................................ 73
A.4 NanoString probe sequence of candidate genes .................................. 74

Appendix B Supplementary material for Chapter 3 .................................. 75
B.1 Discovery and validation workflow of the WRCA panel ..................... 75
B.2 Classification tree of the best performing panel ................................ 76
B.3 List of significant genes comparing expression change from 0h to 6h (PA-positive vs. PA-negative) ......................................................... 77
B.4 Expression levels of the significant genes comparing expression change from 0h to 6h (PA-positive vs. PA-negative) ........................................ 79
List of Tables

Table 2.1 Clinical and demographic characteristics of the individuals in discovery set. ............ 16
Table 2.2 Clinical and demographic characteristics of the individuals in validation set............. 17
Table 2.3 Clinical and demographic characteristics of the individuals in ELISA set. ............... 18
Table 2.4 Differential expression of ERs and DRs in the discovery set at pre-challenge (RNA- Seq). ................................................................................................................................. 24
Table 2.5 Differential expression of ERs and DRs in the discovery set at pre-challenge (NanoString). ................................................................................................................................. 25
Table 2.6 Differential expression of ERs and DRs in the validation set at pre-challenge (NanoString). ................................................................................................................................. 26
Table 3.1 Clinical and demographic characteristics of the individuals in discovery set. ............ 35
Table 3.2 Clinical and demographic characteristics of the individuals in validation set............ 36
List of Figures

Figure 1.1 Early and late phase response of allergic asthma. .......................................................... 4
Figure 1.2 Workflow of a general NanoString gene expression assay. ................................................. 9
Figure 1.3 Thesis Summary. .............................................................................................................. 10
Figure 2.1 The complement and coagulation pathway. .................................................................... 12
Figure 2.2 NanoString nCounter Elements tag-target complex. .................................................... 20
Figure 2.3 Expression levels of gene transcript isoforms that are differentially expressed between ERs and DRs at pre-challenge (RNA-Seq). ................................................................. 23
Figure 2.4 Gene expression level of ERs and DRs in the discovery set at pre-challenge (NanoString). .......................................................................................................................... 25
Figure 2.5 Gene expression level of ERs and DRs in the validation set at pre-challenge (NanoString). ........................................................................................................................................ 27
Figure 2.6 C3a concentration in ERs, DRs and non-asthmatic controls at pre-challenge. ......... 28
Figure 2.7 C5a concentration in ERs, DRs and non-asthmatic controls at pre-challenge. ......... 29
Figure 3.1 Percentage change in FEV₁ in discovery set post methacholine challenge (Day 1). .. 39
Figure 3.2 Percentage change in FEV₁ in discovery set post PA challenge (Day 2). ................. 40
Figure 3.3 Relative proportion (%) of leukocyte subtypes (discovery set) following Day 1 methacholine challenge .................................................................................................................. 42
Figure 3.4 Relative proportion (%) of leukocyte subtypes (discovery set) following Day 2 PA challenge. ........................................................................................................................................... 43
Figure 3.5 AUC performance of all the tested panels in discovery set ............................................. 44
Figure 3.6 AUC performance of the reported WRCA biomarker panel (solid line) and the AUC performance after the reshuffling of phenotypic labels (dashed line) in discovery set. ............. 45

Figure 3.7 Classification of individuals in the discovery set based on log$_2$ expression of $MAP2K2$ and $MAPKAPK2$. ................................................................. 47

Figure 3.8 Probability score of individuals in the discovery set............................................. 48

Figure 3.9 Expression levels of $MAP2K2$ and $MAPKAPK2$ during the 6-hour PA challenge. .... 49

Figure 3.10 Classification of individuals in the validation set based on log$_2$ expression of $MAP2K2$ and $MAPKAPK2$. ................................................................. 50

Figure 3.11 Probability score of individuals in the validation set............................................. 51
## List of Symbols and Abbreviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>alpha</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the receiver operating characteristic curve</td>
</tr>
<tr>
<td>BAL</td>
<td>bronchoalveolar lavage</td>
</tr>
<tr>
<td>BAM</td>
<td>binary compressed sequence alignment map (SAM) format</td>
</tr>
<tr>
<td>BH-FDR</td>
<td>Benjamini-Hochberg false discovery rate</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>CBC/diffs</td>
<td>complete blood cell counts and differentials</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary deoxyribonucleic acid (DNA)</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DR</td>
<td>dual responder</td>
</tr>
<tr>
<td>EAR</td>
<td>early asthmatic response</td>
</tr>
<tr>
<td>ELISA</td>
<td>enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopedia of DNA Elements</td>
</tr>
<tr>
<td>ER</td>
<td>early responder</td>
</tr>
<tr>
<td>ERCC</td>
<td>External RNA Control Consortium</td>
</tr>
<tr>
<td>ERK</td>
<td>extracellular signal-regulated protein kinase</td>
</tr>
<tr>
<td>FEV$_1$</td>
<td>forced expiratory volume in 1 second</td>
</tr>
<tr>
<td>FVC</td>
<td>force vital capacity</td>
</tr>
<tr>
<td>FOV</td>
<td>field of view</td>
</tr>
<tr>
<td>GATA</td>
<td>GATA-binding protein</td>
</tr>
<tr>
<td>GTF</td>
<td>gene transfer format</td>
</tr>
<tr>
<td>HDM</td>
<td>house dust mite</td>
</tr>
<tr>
<td>IgE</td>
<td>immunoglobulin E</td>
</tr>
<tr>
<td>iTRAQ</td>
<td>isobaric tag for relative and absolute quantitation</td>
</tr>
<tr>
<td>JNK</td>
<td>c-Jun N-terminal kinase</td>
</tr>
<tr>
<td>LAR</td>
<td>late asthmatic response</td>
</tr>
<tr>
<td>limma</td>
<td>linear models for microarray and RNA-Seq data</td>
</tr>
<tr>
<td>LOESS</td>
<td>locally weighted polynomial regression</td>
</tr>
<tr>
<td>LOOCV</td>
<td>leave-one-out cross-validation</td>
</tr>
<tr>
<td>MAC</td>
<td>membrane attack complex</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>MAPK</td>
<td>mitogen-activated protein kinase</td>
</tr>
<tr>
<td>MBL</td>
<td>mannose binding lectin</td>
</tr>
<tr>
<td>PA</td>
<td>plactivic acid</td>
</tr>
<tr>
<td>PAI-1</td>
<td>plasminogen activator inhibitor-1</td>
</tr>
<tr>
<td>PAMPs</td>
<td>pathogen-associated molecular patterns</td>
</tr>
<tr>
<td>PC&lt;sub&gt;20&lt;/sub&gt;</td>
<td>provocative concentration of methacholine that causes a 20% drop in FEV&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td>QC</td>
<td>quality control</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>RAM</td>
<td>random access memory</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA-sequencing</td>
</tr>
<tr>
<td>RSEM</td>
<td>RNA-Seq by Expectation-Maximization</td>
</tr>
<tr>
<td>SAM</td>
<td>sequence alignment map</td>
</tr>
<tr>
<td>STAR</td>
<td>Spliced Transcripts Alignment to a Reference</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TF</td>
<td>tissue factor</td>
</tr>
<tr>
<td>Th</td>
<td>T helper</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California, Santa Cruz</td>
</tr>
<tr>
<td>WRCA</td>
<td>western red cedar asthma</td>
</tr>
</tbody>
</table>
Acknowledgements

First, and most of all, I would like to thank my supervisor Dr. Scott Tebbutt who supported me and helped me a lot during my graduate studies. He led me through my master’s studies all the way from an undergraduate student who knew nothing about research to a skillful researcher and provided me with many opportunities to learn. I would like to thank my co-supervisor Dr. Edward Conway who also supported me and gave me many valuable suggestions on both my research and my career path. I also want to thank my supervisor committee Dr. Gabriela Cohen Freue and Dr. Jayachandran Kizhakkedathu for their kind support and great feedbacks. I want to thank our collaborator Dr. Christopher Carlsten, too, for his professional insight towards phenotyping the study patients and his input on the research project.

I would like to acknowledge the financial support from Mitacs, AllerGen NCE, the BC Lung Association, WorkSafe BC and UBC Faculty of Medicine. I would also like to thank Dr. Carlsten’s group at Chan-Yeung Centre for Occupational and Environmental Respiratory Disease and the research staff of AllerGen Clinical Investigator Collaborative in recruiting the research participants and collecting the samples. The studies could not be possible without their excellent work.

I would like to express my sincere thanks to my colleagues and mentors, Dr. Amrit Singh and Mr. Casey Shannon for their great help on statistical and computational data analysis. As a student from a non-statistical background, these research projects could not be accomplished without their patient teaching. I also want to thank my colleagues, Mr. Young Woong Kim, Dr. Emilie Lameignere, Ms. Linnette Ocariza and Mr. Victor Lei, for teaching me experimental laboratory techniques. Thank you to all my previous and current lab members, Mr. Luka Culbrik,
Ms. Amreen Toor, Ms. Jaclyn Rollins, Dr. Tara Fernandez, Dr. Piyushkumar Kapopara, Ms. Jenny Huang and Dr. Houra Loghmani for your generous help.

Lastly, I would like to thank my parents and my dear friend Xuan Zhang for their love and support. Thank you!
To My Parents.
Chapter 1: Introduction

1.1 Thesis overview

Asthma is a chronic inflammatory disorder characterized by airway obstruction. Both allergic asthma and occupational asthma are common types of asthma and greatly diminish patients’ quality of life. This thesis includes two goals: 1) to investigate the relationship between mild allergic asthma and the complement system, and 2) to develop a blood-based biomarker panel to help diagnose western red cedar asthma (WRCA), which is one of the most common forms of occupational asthma in British Columbia.

Upon allergen inhalation, while most mildly allergic individuals develop an early asthmatic response (EAR), approximately 50% of the individuals develop both an EAR and a late asthmatic response (LAR) \(^1\). Individuals who only present with an isolated EAR are classified as early responders (ERs) whereas individuals who present with both an EAR and a LAR are classified as dual responders (DRs). The complement system is a key component of innate immunity and plays an important role in allergic asthma. The first part of the thesis (Chapter 2) was designed to identify complement molecules (genes and proteins) that were differentially expressed in ERs and DRs. Identification of such molecules may help understand the underlying mechanism of the LAR and reveal new therapeutic targets.

The second part (Chapter 3) of the thesis focused on biomarker development of WRCA, which is the most common form of occupational asthma in the Pacific Northwest region of North America, including British Columbia \(^2-^4\). I developed a blood-based biomarker panel that can discriminate patients with WRCA from those suffering from asthma due to other causes, at
baseline without having the patients to go through the inhalational challenges, which may greatly improve the current diagnosis of the disease.

In addition, Chapter 4 introduces an automated ribonucleic acid-sequencing (RNA-Seq) data processing pipeline using “Snakemake” and describes how it can be used towards file management. RNA-Seq data quantification is complicated by a large number of files generated during the process. Having such an automated pipeline greatly reduces the complexity of the process, saves time and helps with project management.

1.2 Allergic asthma

1.2.1 Introduction

Asthma is a chronic inflammatory disorder characterized by airway obstruction and hyperresponsiveness. It affects more than three million Canadians (~10%) but remains poorly understood due to its complexity and heterogeneity. Among these individuals with asthma, 75% belong to the allergic type. Although there has been significant improvement in asthma diagnosis and clinical practice guidelines over the past decade, 50% of Canadians with asthma remain uncontrolled. Symptoms of allergic asthma include coughing, wheezing, shortness of breath and chest tightness.

1.2.2 Clinical diagnosis

Medical history is critically important in the initial diagnosis of asthma. Spirometry is then used to confirm the diagnosis by measuring FEV$_1$ (forced expiratory volume in 1 second) and FEV$_1$/FVC (force vital capacity) ratio. This also provides a quantitative measure to follow. However, patients with normal spirometry results can also present with symptoms of asthma. In
such cases, bronchial provocation tests are used to confirm airway hyperresponsiveness. Methacholine is inhaled in doubling concentrations until the patient has a decrease of 20% in FEV$_1$ (PC$_{20}$). A concentration of less than 16 mg/L is used as an indicator of airway hyperresponsiveness. The lower the concentration, the greater the severity of airway hyperresponsiveness. Besides spirometry, physical examination of wheezing on auscultation can also be used to confirm airflow limitation. Skin prick tests are commonly used to confirm an atopic condition.

1.2.3 Biphasic responses

Both an EAR and a LAR are present in allergic asthma. The EAR occurs several minutes after inhalation of the allergen and can be easily reversed by using bronchodilators. In contrast, the LAR develops 2-6 hours after inhalation and is hard to reverse even with corticosteroids (Figure 1.1). While the majority of individuals with asthma have an EAR, only 50% of individuals develop LAR$^1$. Based on their different responses to the allergen, individuals with an isolated EAR are classified as ERs whereas individuals with both an EAR and a LAR are classified as DRs. Understanding the molecular mechanisms underlying the differences would allow us to better understand disease heterogeneity and to develop new diagnostic tools and treatments.
1.2.4 Disease mechanisms

Over the past 2-3 decades, it has become generally accepted that the allergic response of patients with asthma is associated with T helper type 2 (Th2) immune processes. Upon inhalation of allergen, elevated levels of Th2 cells release cytokines such as IL-4, IL-5, IL-9 and IL-13, which promote inflammation and immunoglobulin E (IgE) production by mast cells. IgE, in turn, triggers mast cell degranulation and the release of inflammatory mediators such as histamine. These mediators lead to the immediate EAR. The mechanism which results in the chronic LAR is not clear at the moment. Most studies believe that the LAR is related to the recruitment of eosinophils and neutrophils and that the structural and functional characteristics of the LAR of
allergic asthma provide the link between the acute immunoglobulin E response and chronic persistent asthma.

1.3 Occupational asthma

1.3.1 Western red cedar asthma

Western red cedar asthma (WRCA) is the most common form of occupational asthma in the Pacific Northwest region of North America including British Columbia and affects approximately 5% of exposed workers. The majority of the affected individuals are sawmill workers, shingle mill workers, carpenters and construction workers. With continuous exposure of approximately 3 years, symptoms including cough, wheeze and dyspnea, start to develop. Initially, the symptoms occur after working hours and the affected individual frequently awakens at night with cough, wheeze and chest tightness. The symptoms tend to appear at work if the exposure continues. Some of the patients also develop persistent breathlessness after several weeks or months. The symptoms show improvement during weekends and holidays but the recovery usually takes longer as the disease progresses. There is an assumption that the symptoms will disappear after removal from the exposure; however, as shown in previous studies, half of the patients continue to have recurrent asthma attacks and require regular medications for symptomatic relief. After years of exposure, patients usually have persistent airflow obstruction and respiratory impairment. Patients with a shorter duration of exposure prior to diagnosis, show better recovery, indicating that early diagnosis and exposure removal is essential for treating the disease and improving health-related quality of life.
1.3.2 Clinical diagnosis

Diagnostic testing for WRCA has not changed over the past several decades and involves multiple inhalational bronchial challenges\(^4,13,14\). The patient needs to stay away from work and stay in the clinic for at least 2 days to confirm the diagnosis. Confirmation of the diagnosis of WRCA requires that the patient experiences a reduction in lung function (i.e. FEV\(_1\)) in response to inhalation of plicatic acid (PA), the disease-causing molecule isolated from western red cedar wood dust. Methacholine challenge tests are also performed to confirm airway hyperresponsiveness. A detailed interview of the patient’s past and current occupation and health status is also required for a diagnosis. Therefore, the whole process can be time-consuming, complicated, expensive, and also may cause discomfort to the patients. Adiponectin has been shown in a previous study to be a potential WRCA biomarker after PA challenge; however, there are no known biomarkers that can predict WRCA at baseline without challenges\(^{15}\).

1.3.3 Disease mechanisms

Three types of asthmatic reactions to PA inhalation are observed in WRCA: isolated EAR, isolated LAR and dual responses. Isolated LAR and dual responses are very common in WRCA and account for more than 90% of the disease population while the isolated EAR accounts for less than 10% of the patients\(^4\). The mechanisms underlying WRCA are not yet well understood, but both immunological and non-immunological mechanisms are likely involved. Specific IgE antibodies to the crude red cedar dust have previously been detected in approximately 40% of patients with WRCA, and were not found in healthy controls or individuals who have asthma but did not respond to PA\(^{16}\). However, the role of IgE antibodies is still not clear since the antibodies were not found in all the patients with WRCA and they were
not only observed in the early responders and dual responders but also observed in the isolated late responders\textsuperscript{3}. On the other hand, as a low molecular weight molecule, PA may act as a hapten to combine with other proteins that are found in the airways to form a complete, functional antigen. This complete antigen may then cross-link with specific IgE to trigger the allergic reaction seen in WRCA\textsuperscript{17}. Besides the immunological mechanisms, previous studies have shown that upon PA exposure, there was direct increased histamine release from human basophils or mast cells in some of the individuals with WRCA in comparison to healthy controls and patients with asthma unrelated to WRCA (non-WRCA)\textsuperscript{3,18}. Activation of the complement system via the classical pathway was also observed in the patients with WRCA\textsuperscript{19}. Proinflammatory molecules such as C3a and C5a generated during complement system activation could also induce histamine release from human basophils and mast cells. In addition, activation of the complement system could result in increased vascular permeability\textsuperscript{20}, which is an important feature of the late phase response. Moreover, upon PA inhalational challenge, there was an increase in sputum eosinophils among patients with WRCA at 6 hours and 24 hours post challenge and this was inversely correlated with the drop of FEV\textsubscript{1} at the 6-hour time point\textsuperscript{21,22}. This suggests that the late phase response occurring in most patients with WRCA may be associated with elevated levels of eosinophils in the airway. Levels of exhaled NO were also increased at 24 hours post PA challenge and the levels of exhaled NO during methacholine challenge and PA challenge were correlated with the level of eosinophilia\textsuperscript{23,24}. Loss of epithelial cells during the late phase response\textsuperscript{25} and increased numbers of T-lymphocytes in the bronchial mucosa\textsuperscript{26} were also observed in individuals with WRCA.
1.4 High throughput molecular technologies

1.4.1 RNA-Seq

Initially, gene expression studies were performed by low-throughput methods such as northern blots and quantitative polymerase chain reaction (qPCR) \(^{27}\). These methods are limited to measuring only one transcript at a time. Over the last two decades, advanced technologies have enabled us to measure gene expression more globally (transcriptomics). Before RNA-Seq technology was developed, transcriptomic studies largely relied on hybridization-based microarrays. Microarrays can simultaneously quantify thousands of transcripts at a relatively low cost; however, several limitations are associated with microarrays. For example, there are cross-hybridization artifacts in the analysis of highly similar sequences and the ability to quantify low or highly expressed genes is very limited \(^{27}\). Compared to microarray, RNA-Seq is more detailed in gene expression quantification. It not only performs better in quantifying low and highly abundant gene transcripts but also enables us to discover new sequences that are not yet annotated. The preparation of RNA-Seq includes the following steps: first, RNA is extracted from biological materials such as cells or tissues; next, RNA is converted to complementary deoxyribonucleic acid (cDNA) by reverse transcription; finally, after PCR amplification, a library is constructed and ready for sequencing \(^{27}\).

1.4.2 NanoString technologies

In comparison to RNA-Seq which has a long work-up (such as cDNA amplification and library construction), the NanoString platform is fully automated and has higher sensitivity, precision, reproducibility and a lower background signal. Also, the NanoString assay is a non-enzymatic assay and inexpensive to implement in a clinical setting with data being generated in
less than 24 hours. The workflow of the NanoString assay is shown in Figure 1.2. The assay consists of both capture and reporter probes. The reporter probe is fluorescently labeled with a unique colour code for each target transcript. In the assay, the target transcripts bind together with reporter probes and capture probes to form target-probe complexes by hybridization at 67°C for 18-24 hours. Then, the hybridized complexes are immobilized and aligned on a glass cartridge. At the final step, the cartridge is scanned in a digital analyzer and each transcript-unique colour code is quantified for its abundance by counting.

![Figure 1.2 Workflow of a general NanoString gene expression assay.](image)

1.5 Thesis Summary

In this thesis, Chapter 2 describes my first project on the relationship between the complement system and the late asthmatic response. I compared both the protein abundance and the gene expression of complement molecules in blood between ERs and DRs at baseline (Figure 1.3). Since the protein abundance did not show any difference and the validation results of gene expression did not corroborate with the discovery results (derived from my analysis of RNA-Seq datasets), I decided to stop this project. However, I was still interested in the late asthmatic response and samples were available to us from another cohort of individuals who were suspected to have western red cedar asthma. Many of these individuals presented a late asthmatic
response, but unfortunately, their phenotypes were not clear and the sample size was small. Therefore, we switched to another research question, which was to identify biomarkers for diagnosing western red cedar asthma (Chapter 3). In this project, I successfully identified and validated a biomarker panel that distinguishes western red cedar asthmatics from non-western red cedar asthmatics (Figure 1.3). Concurrently, to simplify the processing procedures for RNA-Seq data, I developed an automated RNA-Seq data processing pipeline that can be used for any future projects involving RNA-Seq (Figure 1.3). This pipeline is described in Chapter 4 of my thesis.

Figure 1.3 Thesis Summary
Chapter 2: Allergic asthma and the complement system

2.1 Introduction

The complement system is a danger-sensing and host-defense system of innate immunity. Upon recognition of pathogen-associated molecular patterns (PAMPs), the complement system can be activated through three pathways: the classical pathway, the alternative pathway and the lectin pathway. The classical pathway begins with the binding of antigen-antibody immune complexes, PAMPs or apoptotic cells to C1q while the lectin pathway starts with the binding of mannose binding lectin (MBL) or ficolins to carbohydrates on pathogens or apoptotic cells. The alternative pathway is activated through the hydrolysis of the internal thioester bond of C3 to form C3·H₂O. All three types of activation lead to the formation of C3 convertase and converge at the level of C3. C3 can be cleaved into C3a and C3b. C3b participates in the formation of C5 convertase and cleaves C5 into C5a and C5b. C5b, together with C6, C7, C8 and C9, forms the membrane attack complex (MAC), which incorporates into the membranes and induces the lysis of pathogens or cells (Figure 2.1). Anaphylatoxins C3a and C5a generated during this process are well-known proinflammatory molecules which are critical in activating and regulating allergic responses.

The coagulation system is part of the homeostatic process. It is activated through the intrinsic and extrinsic pathways (Figure 2.1). The intrinsic pathway is initiated by contact activation of factor XII while the extrinsic pathway is triggered by tissue factor (TF) which converts factor VII to factor VIIa. Both pathways converge at the level of factor X. The activated form of factor X, factor Xa, converts prothrombin (factor II) into thrombin (factor IIa), which
then catalyzes the cleavage of soluble fibrinogen into fibrin molecules. In the final step, fibrin molecules crosslink with each other to form the fibrin clot with the help of factor XIIIa. There are many crosstalk pathways between the complement and the coagulation system (Figure 2.1)\textsuperscript{29}. For example, C5a is able to enhance blood thrombogenicity by upregulating the expression of TF and plasminogen activator inhibitor-1 (PAI-1) on platelets, mast cells and basophils. Thrombin also contributes to the cleavage of C3 into C3a and C3b, and C5 into C5a and C5b, thus amplifying the activation of the complement system\textsuperscript{29}.

The complement and coagulation systems play a complex role in allergic asthma. Many studies have reported observing complement activation during allergic responses. Nagata and Glovsky showed that extracts of aeroallergens, such as house dust mite (HDM), \textit{Aspergillus
*fumigatus* and perennial ryegrass can induce *in vitro* production of anaphylatoxins C3a and C5a 

In a segmental allergen challenge model of individuals with asthma, levels of both C3a and C5a were increased in bronchoalveolar lavage (BAL) after 24 hours while the levels of C3a and C5a only showed minor increases in healthy individuals. There was also evidence in genome-wide association studies showing linkage of asthma and chromosomal regions coding for C5 and its receptor. In animal studies, upon allergen exposure, C5-deficient mice were more prone to the development of airway hyperresponsiveness and pulmonary inflammation compared to wild-type mice. Further investigation showed that blocking C5a receptor signaling prior to allergen sensitization induced Th2-biased immune responses; however, blockade of C5a receptor in already sensitized mice suppressed airway inflammation and hyperresponsiveness.

### 2.2 Hypothesis

Although it is generally accepted that the complement and coagulation systems participate in allergic responses of asthma, it is not clear whether it affects the late phase response of asthma or not. There are previous studies showing that C3 may regulate the late asthmatic response using mouse models. Using a human allergen inhalation challenge model, our preliminary data in microarrays and iTRAQ (isobaric tag for relative and absolute quantitation) proteomics (see appendix A.1-2 for details) suggested that the expression of complement molecules may be associated with the LAR of allergic asthma. We hypothesized that complement molecules (genes and proteins) are differentially abundant in ERs and DRs of allergic asthma.
2.3 Methods

2.3.1 Human studies

Upon written informed consent, individuals with mild asthma were recruited into the study at the University of British Columbia, McMaster University and Université Laval as part of the AllerGen Clinical Investigator Collaborative. All participants were non-smokers and had been diagnosed with clinically stable asthma with a baseline FEV$_1$ $\geq$ 70% of the predicted value and PC$_{20}$ < 16 mg/mL. None of the patients had other lung diseases, cardiovascular disease or viral infections at least 4 weeks prior to the challenge. None had used inhaled corticosteroids or other medications for asthma (see Diamant et al. 1 for a complete list of inclusion and exclusion criteria).

2.3.2 Inhalational challenges

Skin prick tests were used prior to inhalational challenges to determine the allergen sensitivity for each individual. Methacholine challenges were performed on Day 1 and Day 3 to confirm airway hyperresponsiveness and the allergen-induced shift of PC$_{20}$ ([PC$_{20}$]$_{pre}$/[PC$_{20}$]$_{post}$). Allergen challenge was performed on Day 2, using allergen extracts that were pre-determined by skin prick tests. Both methacholine and allergen extracts were given in doubling doses until a drop of 20% in FEV$_1$ was reached. The initial concentration of allergen extracts was determined based on the results of skin prick tests and Day 1 methacholine challenge. During the inhalational challenges, FEV$_1$ was monitored at regular intervals until 7 hours post-challenge. EAR was defined as having a 20% drop or more in FEV$_1$ between 0 to 2 hours post-allergen challenge while the LAR was defined as having a 15% drop or more between 3 to 7 hours post-allergen challenge. Patients who only presented the EAR were classified as ERs while those who
presented both EAR and LAR were classified as DRs. Individuals who did not fulfil the definition of LAR but demonstrated a drop of 10% or more in FEV$_1$ and an allergen-induced shift ($[PC_{20}]_{	ext{pre}}/[PC_{20}]_{	ext{post}}$) value $\geq 2$ were also classified as DRs.

### 2.3.3 Blood collection

Blood samples were collected at baseline (prior to allergen challenge) and 2 hours post-allergen challenge using standardized protocols. Approximately 5.5 mL of blood was collected at each time point into EDTA tubes (~3ml) and PAXgene Blood RNA tubes (~2.5 mL). Complete blood cell counts and differentials (CBC/diffs) were measured from EDTA blood using an automated hematology analyzer (Cell Dyn 3700 System [Abbott Diagnostics, USA]). Then, the rest of the EDTA blood sample was centrifuged at 500 x g for 10 minutes at room temperature and freshly processed into plasma, buffy coat and erythrocytes. The processed plasma, buffy coat, erythrocytes samples and PAXgene blood samples were stored in -80 °C freezer prior to and after shipment to the Tebbutt laboratory in Vancouver, Canada.

### 2.3.4 Study participants

#### 2.3.4.1 Discovery set

Thirty-six individuals (15 ERs and 21 DRs) were included in the discovery set. Their demographic and clinical characteristics are shown in the following table (Table 2.1). Samples collected from these individuals were used for both RNA-Seq and NanoString analysis.
Table 2.1 Clinical and demographic characteristics of the individuals in discovery set. Statistics are shown as mean ± standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>ERs (n=15)</th>
<th>DRs (n=21)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Female</td>
<td>10 (67%)</td>
<td>14 (67%)</td>
<td></td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>70.2 ± 14.8</td>
<td>73.0 ± 16.3</td>
<td>0.47</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>168.3 ± 8.4</td>
<td>168.4 ± 9.3</td>
<td>0.98</td>
</tr>
<tr>
<td>Age (years)</td>
<td>27.1 ± 8.0</td>
<td>32.0 ± 13.1</td>
<td>0.07</td>
</tr>
<tr>
<td>Baseline FEV$_1$ (L)</td>
<td>3.4 ± 0.8</td>
<td>3.2 ± 0.8</td>
<td>0.44</td>
</tr>
<tr>
<td>% drop in FEV$_1$ during EAR</td>
<td>-35.2 ± 9.9</td>
<td>-35.4 ± 10.0</td>
<td>0.93</td>
</tr>
<tr>
<td>% drop in FEV$_1$ during LAR</td>
<td>-7.5 ± 4.0</td>
<td>-28.7 ± 15.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Allergen induced PC$_{20}$ shift</td>
<td>1.7 ± 1.7</td>
<td>3.4 ± 1.9</td>
<td>0.001</td>
</tr>
<tr>
<td>Allergen (Number of individuals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Fungus</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Grass</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>HDM</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ragweed</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Site (Number of individuals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laval</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>McMaster</td>
<td>5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>UBC</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

2.3.4.2 Validation set

An additional 45 individuals (9 ERs and 36 DRs) were used for validation. Their demographic and clinical characteristics are shown in Table 2.2. Samples collected from these individuals were used for NanoString analysis.
Table 2.2 Clinical and demographic characteristics of the individuals in validation set. Statistics are shown as mean ± standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>ERs (n=9)</th>
<th>DRs (n=36)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Female</td>
<td>5 (56%)</td>
<td>15 (42%)</td>
<td></td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>70.2 ± 11.2</td>
<td>73.3 ± 14.5</td>
<td>0.59</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>167.9 ± 10.3</td>
<td>173.0 ± 8.0</td>
<td>0.13</td>
</tr>
<tr>
<td>Age (years)</td>
<td>33.4 ± 8.2</td>
<td>29.1 ± 10.1</td>
<td>0.24</td>
</tr>
<tr>
<td>Baseline FEV(_1) (L)</td>
<td>3.0 ± 0.5</td>
<td>3.5 ± 0.7</td>
<td>0.048</td>
</tr>
<tr>
<td>% drop in FEV(_1) during EAR</td>
<td>-30.6 ± 10.3</td>
<td>-34.9 ± 9.0</td>
<td>0.22</td>
</tr>
<tr>
<td>% drop in FEV(_1) during LAR</td>
<td>-6.4 ± 4.5</td>
<td>-22.5 ± 8.7</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Allergen induced PC(_{20}) shift</td>
<td>1.2 ± 0.8</td>
<td>3.0 ± 1.8</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Allergen (Number of individuals)</th>
<th>Birch</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cat</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Fungus</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Grass</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>HDM</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Horse</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Ragweed</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Site (Number of individuals)</td>
<td>Laval</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>McMaster</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>UBC</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

2.3.5 ELISA set

Samples collected from 49 individuals (18 ERs, 25 DRs and 6 non-asthmatics) were used for ELISA (enzyme-linked immunosorbent assay) quantification of C3a and C5a levels. The demographic and clinical characteristics of these individuals are shown in Table 2.3. Of the 49 individuals, 31 individuals (14 ERs and 17DRs) overlapped with the discovery set.
Table 2.3 Clinical and demographic characteristics of the individuals in ELISA set. Statistics are shown as mean ± standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>ERs (n=18)</th>
<th>DRs (n=25)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Female</td>
<td>12 (67%)</td>
<td>18 (72%)</td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>28.4 ± 8.2</td>
<td>31.2 ± 12.2</td>
<td>0.50</td>
</tr>
<tr>
<td>Baseline FEV₁ (L)</td>
<td>3.3 ± 0.8</td>
<td>3.2 ± 0.7</td>
<td>0.80</td>
</tr>
<tr>
<td>% drop in FEV₁ during EAR</td>
<td>-34.5 ± 10.3</td>
<td>-37.1 ± 9.0</td>
<td>0.43</td>
</tr>
<tr>
<td>% drop in FEV₁ during LAR</td>
<td>-6.7 ± 4.2</td>
<td>-28.8 ± 12.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Allergen induced PC₂₀ shift</td>
<td>1.4 ± 1.4</td>
<td>3.0 ± 1.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Allergen (Number of individuals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cat</td>
<td>12</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Fungus</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Grass</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>HDM</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ragweed</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Site (Number of individuals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laval</td>
<td>13</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>McMaster</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>UBC</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

2.3.6 Experimental techniques

2.3.6.1 RNA extraction

RNA in blood was extracted from 5 mL of the PAXgene blood samples using the PAXgene Blood miRNA Kit (PreAnalytiX, Switzerland). The RNA concentration and quality was assessed using a NanoDrop 8000 Spectrophotometer (Thermo Scientific, USA) and Agilent 2100 Bioanalyzer system (Agilent Technologies, USA). A RIN (RNA integrity number) of 7.0 was used as the threshold for inclusion.
2.3.6.2 RNA-Seq

The extracted RNA samples in the discovery set were sent to Génome Québec (Centre d’Innovation Genome Quebec et Université McGill) for sequencing. The RNA quality was assessed again by Génome Québec prior to sequencing. External RNA Control Consortium (ERCC) spike-in controls (92 sequences) were added to all samples. rRNA/globin-depleted stranded cDNA libraries were constructed and sequenced using an Illumina HiSeq™ 2000 sequencing system as 100 base pairs (bp) paired-end reads.

2.3.6.3 NanoString nCounter Elements assay

100 ng of the purified total RNA was used in a custom NanoString nCounter Elements™ assay (NanoString Technologies, US) to simultaneously quantify the expression of 166 selected transcripts, including 8 candidates from the complement and coagulation pathways. The candidates were selected based on the analysis of RNA-Seq data. 100 bases of oligonucleotide probes were designed for all selected transcripts using NanoString bioinformatics web-based tools by which users were allowed to select specific probe sequences for transcripts of interest. The NanoString nCounter Elements technology™ consists of an nCounter Elements TagSet (capture and reporter probes) and target-specific oligonucleotide probe pairs (probe A & B) (Figure 2.2). The reporter probe is fluorescently labeled with a unique colour code (sequence of six colour spots) for each target transcript sequence. In the assay, the purified RNA sample was mixed with the specific oligonucleotide probe pairs, the 50-nucleotide-sized reporter probes and the capture probes of the target transcripts for hybridization at 67°C over 16 hours to form the tag-target complex (Figure 2.2). During hybridization, the reporter and the capture tags bound to probe A and B which further bound to the target transcript sequence. Then, the hybridized
samples were processed with a “high sensitivity protocol” on the automated PrepStation for 3 hours to immobilize the tag-target complex to the glass cartridge. The transcript counts were acquired from the GEN2 Digital Analyzer after a 2.5-hour scan.

![Diagram of NanoString nCounter Elements tag-target complex.](image)

**Figure 2.2 NanoString nCounter Elements tag-target complex.**

### 2.3.6.4 ELISA

Baseline blood plasma (prior to allergen inhalation challenge) levels of C3a and C5a were measured by specific ELISAs. The assays were performed in the Conway Laboratory (Centre for Blood Research, Vancouver) using Quidel MicroVue™ C3a Plus EIA Kits and Quidel MicroVue™ C5a Plus EIA Kits, respectively, according to the manufacturer's instructions.

### 2.3.7 Data analysis

#### 2.3.7.1 RNA-Seq processing

All raw RNA-Seq FASTQ files were quality controlled using FastQC standards. The first 5 bases and the last 25 bases were trimmed from the 100 bp paired-end reads using Seqtk (version 1.0). The University of California, Santa Cruz (UCSC) GTF file and 2013 human reference genome (GRCh38 build) were used to build a reference genome sequence for RSEM.
The sequencing data were aligned using Bowtie2 aligner (version 2.2.4) and quantified at both gene and gene-isoform levels using RSEM (version 1.2.19).

The gene and isoform counts were then normalized to log$_2$ counts per million using the following equation from the voom() function of limma:

$$X_{\text{norm}} = \log_2\left(\frac{X_{\text{counts}} + 0.5}{(\text{lib.size} + 1) \times 10^6}\right)$$

where $X_{\text{counts}}$ is a $p$ (variables) by $n$ (samples) matrix of the count data and lib.size is the sum of all counts in each sample.

### 2.3.7.2 NanoString data processing

The NanoString data were normalized and analyzed using the statistical computing environment, R (version 3.2.4). Firstly, the data were assessed for all the recommended quality control metrics including field of view, binding density, positive spike-ins and background signals (see appendix A.3 for details). The Elements assay uses six levels of positive spike-ins: 128fM, 32fM, 8fM, 2fM, 0.5fM and 0.125fM and 13 housekeeping genes. The first 5 levels of positive spike-ins were used for positive control normalization (as recommended by NanoString) to eliminate assay-to-assay variability and the top 5 housekeeping genes with the lowest coefficient of variations were used for house-keeping normalization to correct for the biological variations across the samples. Since there was no obvious batch effect between the data of the discovery set and the data of the validation set, the data of the two sets were normalized together by a positive control normalization followed by a housekeeping normalization.
2.3.7.3 **Differential expression analysis**

Differential expression analysis was applied to the RNA-Seq data of 76 genes involved in the complement and coagulation systems to identify potential candidate genes, and to the NanoString data to validate the candidate genes, using linear models for microarray and RNA-Seq data (limma)\(^43\). Limma shrinks the variance of each variable towards a common value by using a moderation factor, which reduces the number of false positives observed in studies with small sample sizes\(^44\). The Benjamini-Hochberg method was used to control the false discovery rate (BH-FDR)\(^45\) and to adjust for multiple hypotheses testing.

2.3.7.4 **Hypothesis testing on ELISA data**

The absorbance reading at 450 nm of standard dilutions were used to build the standard curves and the C3a and C5a concentrations were calculated from the standard curves. Shapiro-Wilk normality tests were used to assess the normality of the data. Non-parametric Kruskal-Wallis H tests were used to compare the concentrations of C3a and C5a among ERs, DRs and non-asthmatic individuals.

2.4 **Results**

2.4.1 **Transcriptomic difference between ERs and DRs (RNA-Seq)**

Samples from patients in the discovery set were used for RNA-Seq data analyses. Differential expression analysis was performed to compare the difference between ERs and DRs. For discovery purposes, the eight top-ranked gene isoforms were selected as candidates and designed into a custom NanoString nCounter\(^\text{TM}\) Elements assay (see appendix A.4 for gene
names and probe sequences). The log$_2$ fold change, expression levels and $P$-values, BH-FDRs of these isoforms are shown in Figure 2.3 and Table 2.4.

**Figure 2.3** Expression levels of gene transcript isoforms that are differentially expressed between ERs and DRs at pre-challenge (RNA-Seq).
Table 2.4 Differential expression of ERs and DRs in the discovery set at pre-challenge (RNA-Seq). *: average log₂ expression of DRs – average log₂ expression of ERs.

<table>
<thead>
<tr>
<th>UCSC ID</th>
<th>Gene Symbol</th>
<th>logFC*</th>
<th>Average Expression</th>
<th>P-Value</th>
<th>BH-FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>uc001muv.4</td>
<td>CD59</td>
<td>-0.346</td>
<td>5.984</td>
<td>0.014</td>
<td>0.287</td>
</tr>
<tr>
<td>uc001yda.1</td>
<td>SERPINA1</td>
<td>0.422</td>
<td>6.011</td>
<td>0.015</td>
<td>0.287</td>
</tr>
<tr>
<td>uc001hfq.4</td>
<td>CD55</td>
<td>-0.408</td>
<td>6.707</td>
<td>0.026</td>
<td>0.287</td>
</tr>
<tr>
<td>uc001hfr.4</td>
<td>CD55</td>
<td>-1.033</td>
<td>1.464</td>
<td>0.028</td>
<td>0.287</td>
</tr>
<tr>
<td>uc001ydc.4</td>
<td>SERPINA1</td>
<td>0.599</td>
<td>2.174</td>
<td>0.030</td>
<td>0.287</td>
</tr>
<tr>
<td>uc001ycz.4</td>
<td>SERPINA1</td>
<td>0.487</td>
<td>4.252</td>
<td>0.035</td>
<td>0.287</td>
</tr>
<tr>
<td>uc002pgi.1</td>
<td>C5AR1</td>
<td>0.347</td>
<td>5.262</td>
<td>0.037</td>
<td>0.287</td>
</tr>
<tr>
<td>uc001muu.4</td>
<td>CD59</td>
<td>1.135</td>
<td>2.811</td>
<td>0.047</td>
<td>0.320</td>
</tr>
<tr>
<td>uc031yse.1</td>
<td>C1R</td>
<td>0.507</td>
<td>-0.094</td>
<td>0.068</td>
<td>0.343</td>
</tr>
<tr>
<td>uc003mwv.3</td>
<td>F13A1</td>
<td>0.382</td>
<td>8.350</td>
<td>0.072</td>
<td>0.343</td>
</tr>
<tr>
<td>uc001hgm.3</td>
<td>CD46</td>
<td>-0.278</td>
<td>7.846</td>
<td>0.074</td>
<td>0.343</td>
</tr>
<tr>
<td>uc001ggg.1</td>
<td>F5</td>
<td>0.298</td>
<td>5.523</td>
<td>0.076</td>
<td>0.343</td>
</tr>
</tbody>
</table>

2.4.2 Technical validation (NanoString Elements)

A technical validation was applied to the majority (29) of the samples from the discovery set using the custom-designed NanoString nCounter™ Elements assay. The differential expression analysis results at pre-challenge are shown in Table 2.5 and the expression levels of these genes are shown in Figure 2.4. Using a BH-FDR cutoff of 0.1, only F13A1 was differentially expressed between ERs and DRs. SERPINA1 was trending to significance as indicated by the P-value.
Table 2.5 Differential expression of ERs and DRs in the discovery set at pre-challenge (NanoString). *: average log₂ expression of DRs – average log₂ expression of ERs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>logFC*</th>
<th>Average Expression</th>
<th>t</th>
<th>P-Value</th>
<th>BH-FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>F13A1</td>
<td>0.598</td>
<td>10.740</td>
<td>3.054</td>
<td>0.005</td>
<td>0.036</td>
</tr>
<tr>
<td>SERPINA1</td>
<td>-0.207</td>
<td>13.035</td>
<td>-1.681</td>
<td>0.102</td>
<td>0.410</td>
</tr>
<tr>
<td>CD55</td>
<td>-0.076</td>
<td>5.302</td>
<td>-0.746</td>
<td>0.461</td>
<td>0.952</td>
</tr>
<tr>
<td>C5AR1</td>
<td>-0.065</td>
<td>10.740</td>
<td>-0.721</td>
<td>0.476</td>
<td>0.952</td>
</tr>
<tr>
<td>C3AR1</td>
<td>0.059</td>
<td>8.203</td>
<td>0.314</td>
<td>0.756</td>
<td>0.984</td>
</tr>
<tr>
<td>PLAUR</td>
<td>-0.025</td>
<td>10.168</td>
<td>-0.235</td>
<td>0.815</td>
<td>0.984</td>
</tr>
<tr>
<td>CD46</td>
<td>-0.002</td>
<td>12.125</td>
<td>-0.025</td>
<td>0.980</td>
<td>0.984</td>
</tr>
<tr>
<td>CD59</td>
<td>-0.002</td>
<td>9.062</td>
<td>-0.020</td>
<td>0.984</td>
<td>0.984</td>
</tr>
</tbody>
</table>

Figure 2.4 Gene expression level of ERs and DRs in the discovery set at pre-challenge (NanoString).
2.4.3 Biological validation (NanoString Elements)

As an external biological validation, the same NanoString Elements assay was applied to the independent validation set. The differential expression analysis results at pre-challenge are shown in Table 2.6 and the expression levels of these genes are shown in Figure 2.5. Using a BH-FDR cutoff of 0.1, 4 genes (PLAUR, SERPINA1, CD46 and C5AR1) were differentially expressed between ERs and DRs.

Table 2.6 Differential expression of ERs and DRs in the validation set at pre-challenge (NanoString). *: average log$_2$ expression of DRs – average log$_2$ expression of ERs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>logFC*</th>
<th>Average Expression</th>
<th>t</th>
<th>P-Value</th>
<th>BH-FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLAUR</td>
<td>-0.302</td>
<td>10.207</td>
<td>-3.104</td>
<td>0.003</td>
<td>0.025</td>
</tr>
<tr>
<td>SERPINA1</td>
<td>-0.285</td>
<td>12.956</td>
<td>-2.454</td>
<td>0.018</td>
<td>0.071</td>
</tr>
<tr>
<td>CD46</td>
<td>-0.170</td>
<td>11.960</td>
<td>-2.103</td>
<td>0.041</td>
<td>0.088</td>
</tr>
<tr>
<td>C5AR1</td>
<td>-0.236</td>
<td>10.808</td>
<td>-2.068</td>
<td>0.044</td>
<td>0.088</td>
</tr>
<tr>
<td>F13A1</td>
<td>-0.228</td>
<td>10.590</td>
<td>-1.055</td>
<td>0.297</td>
<td>0.457</td>
</tr>
<tr>
<td>C3AR1</td>
<td>-0.156</td>
<td>8.099</td>
<td>-0.954</td>
<td>0.345</td>
<td>0.457</td>
</tr>
<tr>
<td>CD55</td>
<td>0.104</td>
<td>5.188</td>
<td>0.850</td>
<td>0.400</td>
<td>0.457</td>
</tr>
<tr>
<td>CD59</td>
<td>0.021</td>
<td>8.921</td>
<td>0.193</td>
<td>0.848</td>
<td>0.848</td>
</tr>
</tbody>
</table>
2.4.4 Differential abundance of C3a and C5a

C3a and C5a protein levels were measured by ELISA. Non-parametric Kruskal-Wallis H tests was used to compare C3a and C5a levels among ERs, DRs and non-asthmatic controls. There was a significant difference in C3a levels among ERs, DRs and non-asthmatic controls ($P = 0.02$) (Figure 2.6) while there was no difference in the C5a levels among the three groups (Figure 2.7) ($P = 0.54$). Post-hoc tests were performed using Dunn’s test $^{46}$ and the FDR was controlled using the Benjamini-Hochberg method (BH-FDR) $^{45}$. C3a levels in both ERs and DRs were significantly higher than non-asthmatic controls (ER vs. non-asthmatics: BH-FDR = 0.01, DR vs. non-asthmatics: BH-FDR = 0.03) (Figure 2.6).
Figure 2.6 C3a concentration in ERs, DRs and non-asthmatic controls at pre-challenge.
2.5 Discussion

Our NanoString results revealed *SERPINA1* as a gene of interest which may be associated with the late phase response of allergic asthma. Although the differential expression of *SERPINA1* between ERs and DRs was not significant in the discovery set, it was trending to significance and the direction of change was in the same direction as it was in the validation set. *SERPINA1* encodes a serine protease inhibitor which targets elastase, plasmin, thrombin, trypsin,
chymotrypsin, and plasminogen activator. Abnormal expression of α1-antitrypsin (SERPINA1) may lead to an imbalance between α1-antitrypsin and elastase which forms the extracellular matrix. Degradation of the extracellular matrix is an important characteristic of airway inflammation and remodeling. Both α1-antitrypsin and elastase have been reported to be increased in sputum of patients with asthma 47. Gharib et al. have also shown using both Western blotting and shotgun proteomics that α1-antitrypsin was differentially expressed in sputum when comparing patients with or without asthma 48. Since we hypothesized that DRs may be more prone to severe asthma because of the presence of the chronic LAR, α1-antitrypsin may also play a role in progression to the late phase response of allergic asthma.

C3a is a key molecule of the complement system and our ELISA data confirmed previous literature that it is increase in patients with asthma. C3a is involved in the recruitment and activation of leukocytes and can trigger the release of many allergic mediators, such as leukotrienes, histamine, IL-1, IL-6, and TNF 28. C3a can also induce airway mucus secretion, smooth muscle contraction and increased vascular permeability 49.

Although based on our data, C5a is not associated with the late phase response of asthma, there are other studies showing it may be a novel therapeutic target for allergic asthma. Eculizumab is a monoclonal antibody which prevents the formation of C5a and C5b-9 by binding to C5. Gauvreau et al. have reported that a single infusion of eculizumab 24 hours prior to the allergen challenge shows significant attenuation of the LAR 50. However, in another clinical trial outlined in a review by Monk et al., an oral dose of C5aR antagonist (NGD-2000-1) did not show any improvement in lung function 51.
2.6 Conclusions and future directions

Limitations of the study include the small sample size of ERs, the limited consistency across different platforms (RNA-Seq to NanoString), and the heterogeneity of the phenotypes. Although we tried to match the probe sequences for NanoString to the candidate transcript sequences that we selected from RNA-Seq data, this does not guarantee equivalence. Also, some of the participants who went through the allergen challenges multiple times showed different types of responses each time, which led to difficulties in assigning phenotypic labels. In addition, since many of the complement and coagulation proteins are activated upon cleavage (e.g. C3a, C5a, factor Xa, factor IIa, and etc.), they may be different in their protein levels between ERs and DRs while the differences cannot be detected at the gene level. All these issues might contribute to the corroboration challenges that we observed in this study. Based on the current results, the complement and coagulation systems may be associated with allergic asthma, although the connection is not very strong. Further investigations including the evaluation of protein levels are required.
Chapter 3: Biomarkers of western red cedar asthma

3.1 Introduction

Western red cedar asthma (WRCA) is the most common form of occupational asthma in British Columbia and is caused by increased sensitivity to plicatic acid (PA). The current diagnosis requires multiple bronchial challenges which are time-consuming, expensive and may cause discomfort to patients. This whole process takes at least two days and may require the patients to make several visits to the clinic. This may therefore delay the diagnosis of WRCA and hinder recovery from the disease. So far, there is no known molecular biomarker that can be used to diagnose WRCA without inhalational challenges. Peripheral whole blood is a useful and easily obtainable resource for studying WRCA and the transcriptional changes that occur after methacholine inhalation challenge. Instead of the traditional diagnostic method of WRCA which includes multiple inhalational challenges, a blood test would be ideal to more simply obtain an accurate diagnosis.

3.2 Hypothesis / Rationale

We hypothesized that blood transcriptomic signatures could be developed into biomarker panels which would diagnose WRCA through a simple blood test. Patients with asthma who respond to PA are classified as PA-positive and patients with asthma who do not respond to PA are classified as PA-negative. We aimed to develop a blood-based biomarker panel which distinguishes PA-positive individuals from PA-negative individuals prior to the inhalation challenges. Longer-term, these biomarker molecules might also help to further elucidate the disease mechanism underlying WRCA.
3.3 Methods

3.3.1 Inhalational challenges

Upon written informed consent, individuals with known or suspected WRCA were recruited into the study. All study patients underwent a methacholine challenge (Day 1) followed by a PA challenge (Day 2) at Vancouver General Hospital using standardized protocols. Before the challenges, patients were instructed not to use any inhaled short-acting bronchodilators for 12 hours, long-acting bronchodilators for 24 hours, and inhaled corticosteroids for at least 2 weeks. On Day 1, baseline spirometry and methacholine challenge was performed according to the American Thoracic Society guidelines. On Day 2, baseline spirometry was again obtained. PA was given to the patients in doubling doses (0.625, 1.25, 2.5, 5, and 10 mg/mL) and their FEV₁ was measured after each increase in concentration until a drop of 20% in FEV₁ was observed. On both days, FEV₁ was monitored at regular intervals until 6 hours (20 min, 30 min, 40 min, 60 min and then hourly) after the commencement of inhalational challenge.

Patients with a positive response to methacholine were considered to have asthma and patients with a positive response to PA were diagnosed as having WRCA (PA-positive). According to previous literature, an EAR was defined a drop in FEV₁ of 20% within the first 2 hours post-challenge and a LAR was defined as having a drop in FEV₁ of 15% after 2 hours post-challenge. All PA-positive patients had either an EAR or a LAR, or both. Patients who did not respond to PA within the 6 hours of monitoring were diagnosed as PA-negative.
3.3.2 Blood collection

Peripheral whole blood was collected in EDTA tubes and PAXgene Blood RNA tubes on both days prior to the challenges, and 2 hours and 6 hours post-challenge. After collection, blood samples were transferred to the Tebbutt laboratory at St. Paul’s Hospital for further processing. CBC/diffs were measured using EDTA blood samples to assess for changes in cellular composition. Then, 1 mL of each EDTA blood sample was aliquoted and the rest of the sample was centrifuged at 1200 x g for 10 minutes at room temperature and freshly processed into plasma, buffy coat and erythrocytes. The processed plasma, buffy coat, and erythrocyte samples and PAXgene blood samples were stored in a -80 ºC freezer.

3.3.3 Study participants

3.3.3.1 Discovery set

Seventeen individuals (8 PA-negative and 9 PA-positive) were selected for our discovery biomarker study. The clinical and demographic characteristics of the individuals are shown in Table 3.1. All 3 time-series blood samples collected on Day 2 PA-challenge (baseline 0h, 2h post-challenge and 6h post-challenge) were used in the study. Mann-Whitney U tests were used to compare the clinical and demographic characteristics between the PA-positive group and the PA-negative group.
Table 3.1 Clinical and demographic characteristics of the individuals in discovery set. Statistics are shown as mean ± standard deviations. *: Mann-Whitney U tests comparing PA-positive and PA-negative individuals.

<table>
<thead>
<tr>
<th></th>
<th>PA-Positive (n=9)</th>
<th>PA-Negative (n=8)</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Year)</td>
<td>48.0±12.3</td>
<td>38.8±12.8</td>
<td>0.152</td>
</tr>
<tr>
<td>BMI</td>
<td>28.40±3.70</td>
<td>29.35±6.97</td>
<td>0.962</td>
</tr>
<tr>
<td>Predicted FEV₁ (L)</td>
<td>3.84±0.68</td>
<td>4.15±0.53</td>
<td>0.290</td>
</tr>
<tr>
<td>Predicted FEV₁/FVC</td>
<td>0.90±0.19</td>
<td>0.81±0.02</td>
<td>0.736</td>
</tr>
<tr>
<td>Baseline FEV₁ (L)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>before methacholine challenge</td>
<td>2.91±0.94</td>
<td>3.33±0.79</td>
<td>0.481</td>
</tr>
<tr>
<td>before PA challenge</td>
<td>2.83±0.94</td>
<td>3.18±0.81</td>
<td>0.481</td>
</tr>
<tr>
<td>Current Smoker</td>
<td>n = 0</td>
<td>n = 1</td>
<td></td>
</tr>
</tbody>
</table>

3.3.3.2 Validation Set

An additional 7 PA-positive individuals were used for external validation. Their clinical and demographic characteristics are shown in Table 3.2. Only baseline (0h) blood samples collected on Day 2 PA-challenge were used in the validation study. Kruskal-Wallis H tests were used to compare the clinical and demographic characteristics among PA-positive group and PA-negative group in the discovery set, and PA-positive individuals in the validation set (3-group comparison).
Table 3.2 Clinical and demographic characteristics of the individuals in validation set. Statistics are shown as mean ± standard deviations. *: Kruskal-Wallis H tests comparing PA-positive individuals and PA-negative individuals in the discovery set, and PA-positive individuals in the validation set.

<table>
<thead>
<tr>
<th></th>
<th>PA-Positive (n=7)</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Year)</td>
<td>52.4±11.7</td>
<td>0.122</td>
</tr>
<tr>
<td>BMI</td>
<td>28.86±3.27</td>
<td>0.971</td>
</tr>
<tr>
<td>Predicted FEV₁ (L)</td>
<td>3.86±0.83</td>
<td>0.509</td>
</tr>
<tr>
<td>Predicted FEV₁/FVC</td>
<td>0.95±0.27</td>
<td>0.928</td>
</tr>
<tr>
<td>Baseline FEV₁ (L)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>before methacholine challenge</td>
<td>2.93±0.89</td>
<td>0.618</td>
</tr>
<tr>
<td>before PA challenge</td>
<td>2.74±0.79</td>
<td>0.516</td>
</tr>
<tr>
<td>Current Smoker</td>
<td>n = 0</td>
<td></td>
</tr>
</tbody>
</table>

3.3.4 Experimental techniques

3.3.4.1 Blood RNA preparation

Total RNA was extracted and purified from 5 ml of PAXgene blood samples using the PAXgene Blood miRNA Kit (PreAnalytiX, Switzerland). The RNA concentration and quality was assessed using NanoDrop 8000 Spectrophotometer (Thermo Scientific, USA) and Agilent 2100 Bioanalyzer system (Agilent Technologies, USA). A RIN number of 7 was used as a threshold for sample inclusion.

3.3.4.2 NanoString nCounter Elements assay

100 ng of the purified total RNA was used in the same NanoString nCounter Elements™ assay (NanoString Technologies, US) described in Chapter 2 to simultaneously quantify the expression of 166 transcripts, which are mostly involved in immune processes such as the Th2 pathway. Please see Chapter 2 for details.
3.3.5 Statistical methodologies

3.3.5.1 Data normalization and batch correction

The obtained NanoString data were normalized and analyzed within the statistical computing environment R (version 3.2.4). Firstly, the data were assessed for all the recommended quality control metrics as described in Chapter 2 (Section 2.3.7.2). The first 5 concentrations of positive spike-ins were used for positive control normalization (as recommended by NanoString) to eliminate inter-assay variability and the top 5 housekeeping genes with the lowest coefficient of variations were used for house-keeping normalization to correct for the biological variations across the samples. The data were normalized by a positive control normalization followed by a housekeeping normalization for both the discovery set and the validation set. Since there was a batch effect between the two sets of data, the data from the discovery set and the data from the validation set were normalized separately, and then batch-corrected using Combat R package.

3.3.5.2 Comparisons on CBC/diffs

The relative levels of leukocyte subtypes were compared at each time point and throughout the challenges between the PA-positive group and PA-negative group of the discovery set using non-parametric Mann-Whitney U tests.

3.3.5.3 Biomarker analysis

After normalization, a biomarker discovery analysis was performed to identify biomarker panels that can differentiate PA-positive individuals from PA-negative individuals at baseline
The panels were identified using a biomarker development pipeline, which includes various classification methods such as elastic net, random forests and Support Vector Machines (SVM). The pipeline also incorporated a pathway-directed approach in which a list of genes was pre-selected if they were in the same pathway. Then, these genes (or features) were further selected using various classification algorithms. Area under the receiver operating characteristic curve (AUC) obtained from leave-one-out cross-validation (LOOCV) was used as a criterion to assess the performance of the identified panels. The best performing panel was then tested using individuals from the validation set (see appendix B.1 for a detailed workflow of the discovery and validation procedures).

### 3.3.5.4 Differential expression analysis

Differential expression analysis was performed on all 166 transcripts using limma R package. The Benjamini-Hochberg method was used to control the false discovery rate (BH-FDR) and adjust for multiple hypotheses testing.

### 3.4 Results

#### 3.4.1 Clinical and demographic characteristics

As shown in Table 3.1, there was no significant difference in age, BMI, or baseline FEV₁ (before methacholine challenge and before PA challenge) between the PA-positive group and the PA-negative group in the discovery set using non-parametric Mann-Whitney U tests. There was no significant difference in % change in FEV₁ from baseline at any time point between the two groups post Day 1 methacholine challenge (Figure 3.1). During the course of the Day 2 PA
challenge (Figure 3.2), the PA-positive individuals experienced significant drops in FEV$_1$ compared to the PA-negative individuals at all time points post-challenge ($P < .001$).

Figure 3.1 Percentage change in FEV$_1$ in discovery set post methacholine challenge (Day 1). Statistics are shown as mean ± standard deviations.
Figure 3.2 Percentage change in FEV₁ in discovery set post PA challenge (Day 2). Statistics are shown as mean ± standard deviations.

3.4.2 Comparisons on CBC/diffs

Figures 3.3 and 3.4 show the relative proportions of leukocyte subtypes (neutrophils, lymphocytes, monocytes, eosinophils and basophils from left to right) of individuals in the discovery set following Day 1 methacholine challenge and Day 2 PA challenge, respectively. Two-sample t-tests were used to compare the relative proportions of leukocyte subtypes between PA-positive individuals and PA-negative individuals. There was no significant difference in any of the leukocyte subtypes at each time point following Day 1 methacholine challenge or following Day 2 PA challenge. Paired tests were used to compare the leukocytes subtype
proportion changes between time points, between PA-positive individuals and PA-negative individuals. To achieve this, mean leukocytes subtype proportion differences were first calculated by subtracting paired samples between time points. The mean differences were then compared between PA-positive individuals and PA-negative individuals using two-sample t-tests. During Day 2 PA challenge, neutrophils changed differently from 0h to 6h ($P = 0.046$) and from 2h to 6h ($P = 0.005$) between the PA-positive group and the PA-negative group (Figure 3.4); lymphocytes changed differently from 2h to 6h ($P = 0.023$) between the PA-positive group and the PA-negative group (Figure 3.4); eosinophils changed differently from 0h to 6h ($P = 0.046$) and from 2h to 6h ($P = 0.016$) between the PA-positive group and the PA-negative group (Figure 3.4). Solid lines in the figure represent the locally weighted polynomial regression (LOESS) of the data across the 3 time points and the shaded area represent the 95% confidence interval of the regression model.
Figure 3.3 Relative proportion (%) of leukocyte subtypes (discovery set) following Day 1 methacholine challenge.
3.4.3 Biomarker panel to distinguish PA-positives and PA-negatives

Having demonstrated that there was no difference in the clinical and demographic characteristics between the PA-positive and the PA-negative groups (excepting following PA challenge) in the discovery set, we undertook a biomarker discovery approach to identify a classifier of PA-positive individuals at baseline (see section 3.3.5.3 and appendix B.1 for details). Figure 3.5 shows the cross-validated AUC of all the tested panels.
Figure 3.5 AUC performance of all the tested panels in discovery set. Dotted line indicates an AUC performance of 0.5.

Of all the panels tested, a two-gene panel (MAP2K2 and MAPKAPK2) identified using a MAPK signaling pathway-directed approach with random forests demonstrated the highest AUC performance of 0.847 (95% Confidence Interval: 0.631-1.000) (Figure 3.6) (see section 3.3.5.3 and appendix B.1 for details about the biomarker analysis pipeline). The phenotypic labels (PA-positive or PA-negative) were then reshuffled and the AUC was recalculated 200 times using
LOOCV. The AUC performance (mean ± standard deviation) after reshuffling the phenotypic labels was 0.600 ± 0.095 (Figure 3.6).

Figure 3.6 AUC performance of the reported WRCA biomarker panel (solid line) and the AUC performance after the reshuffling of phenotypic labels (dashed line) in discovery set. Error bars indicate mean sensitivity ± standard deviations. (Reprinted with permission of the American Thoracic Society. Copyright (c) 2017 American Thoracic Society. Cite: Chen Xi Yang, Amrit Singh, Young Woong Kim, Edward M Conway, Christopher Carlsten, Scott J Tebbutt/2017/Diagnosis of Western Red Cedar Asthma Using a Blood-based Gene Expression Biomarker Panel/American Journal of Respiratory and Critical Care Medicine/Epub ahead of print. The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society.)
Random forests are a tree-based ensemble method for machine learning. It combines many predictive tree classifiers where the two nodes of each tree represent the phenotypic classes (in the case of binary classification) and the branches are split based on the value of a certain feature. The features are ranked based on their importance score which reflects their appearing frequencies and abilities to split the phenotypic classes.

The classification was based on the log$_2$ expression of $MAP2K2$ and $MAPKAPK2$ (see appendix B.2 for the classification tree). The classification of individuals in the discovery set and their probability scores are shown in Figure 3.7 and 3.8, respectively. The optimal probability threshold was calculated using the Youden Index. Individuals with a probability score higher than 0.362 were classified as PA-positive.
Figure 3.7 Classification of individuals in the discovery set based on log2 expression of MAP2K2 and MAPKAPK2.
3.4.4 Differential gene expression between PA-positives and PA-negatives

Comparisons of the expression of the genes were made between the PA-positive group and the PA-negative group at baseline (prior to PA challenge) and throughout the course of PA-challenge (expression change from 0h to 6h). At an BH-FDR cutoff of 10%, there were no genes that were significantly differentially expressed between the two groups. However, restricting the
comparisons to the two panel genes (MAPK2 and MAPKAPK2), both MAP2K2 and MAPKAPK2 showed differential expression between PA-positive and PA-negative individuals (BH-FDR = 0.027). The expression levels of MAP2K2 and MAPKAPK2 during the 6-hour PA challenge are shown in Figure 3.9. Thirty-six genes showed significant differences in expression change from 0h to 6h (expression of 6h minus expression of 0h) between the PA-positive group and the PA-negative group using an BH-FDR cutoff of 10% (see appendix B.3-4 for details). Solid lines in the figure represent the LOESS regression of the data across the 3 time points and the shaded area represent the 95% confidence interval of the regression model.

Figure 3.9 Expression levels of MAP2K2 and MAPKAPK2 during the 6-hour PA challenge.

3.4.5 Biomarker panel validation

To validate this panel, 7 additional PA-positive individuals (validation set) were tested. Batch correction across the discovery and validation sets was performed using ComBat$^{55}$ R
package. Batch difference was used as the only variable during correction. The classification model built using original (uncorrected) data from the discovery set was recalibrated after batch correction. Using the recalibrated model and the threshold (0.362) that was previously established using the discovery samples, 6 (86%) of these new individuals were correctly classified as PA-positive. The classification of individuals in the validation set and their probability scores are shown in Figure 3.10 and 3.11, respectively.

Figure 3.10 Classification of individuals in the validation set based on log2 expression of MAP2K2 and MAPKAPK2.
3.5 Discussion

The mitogen-activated protein kinase (MAPK) signaling pathway represents a convergence of many cellular events and participates in many physiological functions. Kinases such as MAPKAPK2 and MAP2K2 in the MAPK signaling pathway play an important role in
activating and regulating inflammatory responses. MAPK pathway activates when stimuli such as growth factor, cell adhesion, cellular and oxidative stress, and inflammatory responses are received. Upon “switching on” the activator protein, the MAP kinases are activated through a phosphorylation cascade, where MAP3Ks (MAP kinase kinase kinases) turns into MAP2Ks (MAP kinase kinases), and then MAPKs (MAP kinases) \(^56\). The MAPK signaling pathway divides into three major groups: extracellular signal-regulated protein kinase (ERK), c-Jun N-terminal kinase (JNK), p38 MAPK. MAPKAPK2 is a downstream molecule of p38 MAPK while MAP2K2 belongs to the ERK group. The p38 MAPK group is activated upon inflammatory responses and is involved in many aspects of asthma such as cytokine release, mast cell migration, monocyte differentiation and neutrophil recruitment \(^57\). It is a key player in activating transcription factor GATA-binding protein (GATA) 3, which further regulates Th2 cell differentiation and Th2 cytokine expression \(^58\). It also contributes to eosinophil differentiation \(^59\) and inhibition of eosinophil apoptosis \(^60\), which may be associated with the elevated levels of these cells observed in WRCA \(^61\). Consistently, in our CBC/diff data, compared to PA-negative individuals, PA-positive individuals showed a greater decrease in the relative abundancy of blood eosinophils from 0h to 6h post-PA challenge, which may suggest a migration of eosinophils from blood to the airway. In addition, previous studies have also shown that inhibitors of p38 MAPK are effective in reducing lung inflammation in animal models of asthma and chronic obstructive pulmonary disease \(^62\). MAP2K2 (or more commonly known as MEK2), is one of the MAP2Ks in the ERK group. Upon activation of the Ras protein, the Raf protein (MAP3K), MEKs (MAP2K) and ERKs (MAPK) are activated in a cascaded fashion. The ERK group is involved in Th2 cell differentiation and cytokine production as well \(^57\). It regulates
airway smooth muscle cell proliferation\textsuperscript{63}, which may further contribute to remodeling and narrowing of the airway.

### 3.6 Conclusions and future directions

Limitations of this study include small sample size (for both discovery and validation) and the functional significance of blood, reflecting changes in lung function. To ensure that the results were not confounded by other variables, we carefully matched the clinical and demographic characteristics of the two groups. In conclusion, we have demonstrated for the first time that peripheral whole blood may have utility as a resource to help diagnose WRCA and that the transcriptional signatures may help to further elucidate the disease mechanisms of WRCA. Further validation of the biomarker panel with larger sets and additional PA-negative individuals is required to confirm its robustness. Since the majority of the study individuals were dual responders and isolated early responders, the panel also needs to be further tested with a more mixed group of PA-positive individuals (including isolated late responders). In addition, although the panel showed adequate classification performance using baseline blood samples prior to Day 2 PA challenge, it is not clear whether Day 1 methacholine challenge affected the classification performance or not. In the future, we would also like to examine this effect using baseline blood samples from Day 1 methacholine challenge.
Chapter 4: Pipeline for RNA-Seq data processing

4.1 Introduction

RNA-Seq is a powerful technology which allows unbiased profiling of the entire transcriptome. RNA-Seq does not rely on transcript-specific probes and therefore, supports quantification of novel transcripts. It is also less affected by background noise and signal saturation compared to microarrays. After sequencing, the obtained raw reads usually require the following processing before they are ready for further analysis: 1) quality control, 2) read alignment, and 3) read summarization. During the whole process, different tools are used and a large number of files are generated. Installation of these tools can be tedious and each tool has its own parameters, input files and output files. An automated pipeline is needed to reduce the complexity of this process and to make the data more reproducible. In this chapter, I developed an automated transcript quantification pipeline which was powered by a workflow management tool called “Snakemake”.

4.2 Methods

4.2.1 Workflow management with Snakemake

The pipeline was implemented as a Snakemake workflow. This workflow is composed of a series of “rules”. Each rule specifies the input and output files and invokes shell commands that create the output files. Dependencies between the rules are determined automatically, and the jobs can be parallelized across cores, or even computer cluster nodes. Since Snakemake only updates output files if the input files are more recent, the workflow can be suspended or resumed at any time, without having to start over. Rules are carried by a Snakefile, which can serve as
a record of the input files, parameters and the programs used in the workflow. The inputs and program parameters can also be stored into a configuration file, which allows further flexibility.

The workflow template repository is available on Github and can be cloned with the following command lines in the terminal:

```
# shell
# clone the repository
git clone https://github.com/PROOF-centre/rnaseq_workflow.git
# move into the repository
cd rnaseq_workflow
# inspect its contents
ls -la
```

The folder contains a Snakefile, where the rules are defined, a config.yaml file, where the inputs and parameters are stored, and a requirements.yaml file, which lists the software used in the workflow. Here I will introduce each rule before all the rules are assembled in the section “All together now!” (Section 4.2.4).

4.2.2 Quality control with FastQC

The first part of the pipeline runs quality control over the downloaded FASTQ files using FastQC. FastQC performs simple checks on the raw sequencing data and produces a HTML-formatted report on the analyses. These analyses can help determine whether the sequencing data has any problems that we should be aware of before doing any further analysis. Interpretations of the diagnostic plots in the report can be found on the FastQC website.

This Snakemake rule takes all the (compressed) FASTQ files as input, passes the files one by one (or in pairs for paired-end reads) to the FastQC program and calls the shell
commands on the input files. After quality evaluation, the output reports are stored in a new folder called “FastQC”, separated with sample names. Running time will depend on the total number of reads (approximately 10 min/FASTQ file for 50M reads). The benchmarks such as running time for each sample can be stored in the “benchmarks” folder.

**Snakemake Rule**

```bash
rule fastQC:
    input:
        lambda wildcards: config["samples"]{wildcards.sample}
    output:
        "fastQC/{sample}/"
    benchmark:
        "benchmarks/fastQC/{sample}.benchmark.txt"
    threads: 8
    params:
        outDir = "fastQC/{sample}/"
    shell:
        "zcat -c {input} | fastqc -t {threads} --outdir {params.outDir} stdin"
```

### 4.2.3 Quantifying transcript abundance with RSEM

The next part of the pipeline quantifies the transcript abundance in each sample using the RSEM (RNA-Seq by Expectation-Maximization) \(^64\) and STAR (Spliced Transcripts Alignment to a Reference) \(^65\) programs. This is done by a two-step process: reads are first aligned to a reference genome sequence by STAR and then summarized both at gene level and isoform level by RSEM. For simplicity, RSEM includes a script — `rsem-calculate-expression` — that performs alignment and summarization in a single command using the parameters recommended by the Encyclopedia of DNA Elements (ENCODEx) consortium. This second part of the data processing pipeline downloads and prepares the reference genome sequence, aligns the reads from each FASTQ file (or, in pairs in the case of paired-end reads) and summarizes them to gene and isoform counts for further analysis.
4.2.3.1 Read alignment

In the first step, reads are aligned to the reference genome sequence by the STAR aligner. STAR is a multi-threaded, high performance alignment program, purpose-built for RNA-Seq. In particular, it greatly improves alignment accuracy for intron-spanning reads. STAR was chosen over alternatives, such as bowtie, mainly because of its flexibility and speed. It achieves a very high alignment rate by using a specialized in-memory reference genome index and this index can be quite large: in the case of the human genome, the system should have > 30GB of random access memory (RAM).

4.2.3.1.1 Generating the STAR index

A reference genome index needs to be built in the first place, so that it can subsequently be loaded into memory prior to carrying out any read alignments. Both the reference genome sequence in a FASTA-formatted file and the gene annotation file in gene transfer format (GTF) will be used in this step. The GTF file is a tab-delimited text file format used to hold information about gene structure. The most current human reference genome and transcriptome sequences, and corresponding annotation files, can be obtained from the GENCODE website. Here I defined a pair of Snakemake rules to download these files from the Sanger Institute FTP servers to the ~/genome/ directory on the system. The download location is specified by the downloadDir parameter in the config.yaml file and can be easily modified by the user. These two rules will be automatically skipped if the ~/genome/ directory contains the downloaded FASTA and GTF files or the reference genome has been built already.
Snakemake Rule

```bash
rule downloadGTF:
  output:
    expand("{downloadDir}gtf/gencode.v25.annotation.gtf",
            downloadDir = config["downloadDir"])
  params:
    downloadDir = config["downloadDir"]
  shell:
    "wget -O -{params.url} | gunzip -c > {output}"
```

Snakemake Rule

```bash
rule downloadFASTA:
  output:
    expand("{downloadDir}fasta/GRCh38.p7.genome.fa", downloadDir = config["downloadDir"])
  params:
    downloadDir = config["downloadDir"]
  shell:
    "wget -O -{params.url} | gunzip -c > {output}"
```

After the FASTA and GTF files are downloaded to the ~/genome/ directory, the next rule invokes the `rsem-prepare-reference` RSEM script, which will call STAR and build a suitable reference genome index.

Snakemake Rule

```bash
rule generateRSEMIndex:
  input:
    gtf = expand("{downloadDir}gtf/gencode.v25.annotation.gtf",
                 downloadDir = config["downloadDir"]),
    fasta = expand("{downloadDir}fasta/GRCh38.p7.genome.fa",
                   downloadDir = config["downloadDir"])
  output:
    expand("{downloadDir}index/rsem/", downloadDir =
            config["downloadDir"])
  shell:
    "rsem-prepare-reference \
```
When the pipeline is invoked for the first time, it calls the `downloadGTF`, `downloadFASTA`, and `generateRSEMIndex` rules to download the GTF annotation and reference genome sequence FASTA files, and generate the necessary STAR genome index. If the genome index already exists, however, the pipeline will skip these rules and immediately start to process the FASTQ files.

### 4.2.3.1.2 Aligning reads

Once the genome index has been built, we are ready to align the reads. STAR will take either a single, or a pair of (compressed) FASTQ files for each sample, take each read or read-pair (sometimes referred to as a fragment) in turn, and find the optimal alignment along the reference genome sequence. Reads (or fragments), along with location of the optimal alignment, are returned and written to a new file, called a BAM file. This is the binary version of a sequence alignment map (SAM) file, a tab-delimited text file standard for sequence alignment data. Since the `rsem-calculate-expression` script of RSEM does read alignment and summarization in a single shell command, no rule was defined specifically for alignment.

### 4.2.3.2 Read summarization

This part of the pipeline calls the `rsem-calculate-expression` script of RSEM to invoke the STAR aligner to perform read alignment and the RSEM program to produce two tab-delimited text files containing the summarized read counts at both the gene level.
and the isoform level. This Snakemake rule passes single, or pairs of (compressed) FASTQ files to STAR (via \texttt{rsem-calculate-expression}) for alignment using the created genome index. Conveniently, STAR can read from compressed FASTQ files directly (specified using \texttt{--star-gzipped-read-file}). Here I chose to discard the large BAM files following quantification with the flag \texttt{--no-bam-output}, but this can be modified in the \texttt{Snakefile} if the BAM files are of interest.

\textbf{Snakemake Rule}

\begin{verbatim}
Snakemake Rule

rule RSEM:
  input:
    fastqs = lambda wildcards: 
      config["samples"][wildcards.sample],
    ref = expand("{downloadDir}index/rsem/", downloadDir = config["downloadDir"])
  output:
    "rsem/{sample}.genes.results",
    "rsem/{sample}.isoforms.results"
  benchmark:
    "benchmarks/rsem/{sample}.benchmark.txt"
  threads: 8
  params:
    readsType = config["readsType"],
    id = "{sample}" 
  shell:
    "rsem-calculate-expression \ 
    \{params.readsType\} \ 
    \--star \ 
    \--star-gzipped-read-file \ 
    -p \{threads\} \ 
    \--no-bam-output \ 
    \{input.fastqs\} \ 
    \{input.ref}GRCh38 \ 
    \{params.id\} & & \
    mv -v *.genes.results *.isoforms.results *.stat/ -t rsem/"
\end{verbatim}

\textbf{4.2.4 All together now!}

Having described all the rules required to go from raw RNA-Seq reads contained in FASTQ files to summarized read counts for a set of annotated genomic features, here I put
together all the defined rules into a single Snakefile that both powers, and serves as a record of the data processing pipeline. I will also describe the config.yaml file -- a mechanism that allows us to customize this pipeline for different analytical requirements.

4.2.4.1 The Snakefile

The Snakefile starts by specifying that the parameters are stored in a configuration file called config.yaml. Besides the rules that have been described in previous sections, we have a top-level rule (rule all) which invokes the other rules based on its dependencies. It also stores the current working environment, including the software installed and their version numbers, into an environment.yaml file for future reference.

```yaml
configfile: "config.yaml"

rule all:
    input:
        expand("fastQC/{sample}/", sample = config["samples"],
    expand("rsem/{sample}.genes.results", sample =
    config["samples"],
    expand("rsem/{sample}.isoforms.results", sample =
    config["samples"])
    shell:
        "conda env export > environment.yaml"

rule downloadGTF:
    output:
        expand("{downloadDir}gtf/gencode.v25.annotation.gtf",
    downloadDir = config["downloadDir"])
    params:
        downloadDir = config["downloadDir"]
    shell:
        "wget -O - {params.url} | gunzip -c > {output}"

rule downloadFASTA:
    output:
        expand("{downloadDir}fasta/GRCh38.p7.genome.fa", downloadDir =
    config["downloadDir"])
```

61
params:
  downloadDir = config["downloadDir"]
shell:
  "wget -O -{params.url} | gunzip -c > {output}"

rule generateRSEMIndex:
  input:
    gtf = expand("{downloadDir}gtf/gencode.v25.annotation.gtf",
    downloadDir = config["downloadDir"])
    fasta = expand("{downloadDir}fasta/GRCh38.p7.genome.fa",
    downloadDir = config["downloadDir"])
  output:
    expand("{downloadDir}index/rsem/", downloadDir =
    config["downloadDir"])
  shell:
    "rsem-prepare-reference \
    -p 8 \ 
    --star \ 
    --gtf {input.gtf} \ 
    {input.fasta} \ 
    {output}GRCh38"

rule fastQC:
  input:
    lambda wildcards: config["samples"][wildcards.sample]
  output:
    "fastQC/{sample}/"
  benchmark:
    "benchmarks/fastQC/{sample}.benchmark.txt"
  threads: 8
  params:
    outDir = "fastQC/{sample}/"
  shell:
    "zcat -c {input} | fastqc -t {threads} --outdir {params.outDir} stdin"

rule RSEM:
  input:
    fastqs = lambda wildcards:
    config["samples"][wildcards.sample],
    ref = expand("{downloadDir}index/rsem/", downloadDir =
    config["downloadDir"])
  output:
    "rsem/{sample}.genes.results",
    "rsem/{sample}.isoforms.results"
  benchmark:
    "benchmarks/rsem/{sample}.benchmark.txt"
  threads: 8
4.2.4.2 The config.yaml file

While some commonly used parameters, such as taking the compressed format of
FASTQ files as input, are specified in the Snakefile, some parameters, like whether to run
STAR or RSEM in paired-end mode, may need to be modified for every new analysis. In the
reported pipeline, the commonly used parameters were directly defined in the Snakefile while
the project-specific parameters were stored in the config.yaml file. Each new project can be
initialized with a static copy of the Snakefile and project-specific parameters can be modified
in the config.yaml file. The config.yaml file is divided into two parts: the first part contains
the parameters passed to RSEM and the STAR aligner (e.g. the path to the genome reference
folder or whether the reads are paired-end reads or not) while the second part lists the samples to
be processed, including sample IDs and paths to their corresponding FASTQ file(s).

4.2.4.3 Running the pipeline

A number of programs need to be installed and placed on the system’s path before
running the pipeline. The following steps describe the way of setting up a suitable environment
on a Linux system, as a regular user without administrative privileges, using Miniconda.

Continue from section 4.2.1 where the template folder has been downloaded,

**1. At the command line, install Miniconda.**

```shell
# download the install script
wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh

# run it
bash Miniconda3-latest-Linux-x86_64.sh

# cleanup
rm Miniconda3-latest-Linux-x86_64.sh
```

**2. Create a named Miniconda environment with the required software.**

```shell
conda env create -n "rnaseq" -f "requirements.yaml"
```

**3. Activate the newly created Miniconda environment.**

```shell
source activate "rnaseq"
```

All software specified in the `requirements.yaml` file should now be available to use at the command line.

**4. Create a `fastq` folder under the project directory.**

```shell
# create a new directory for the FASTQ files
mkdir fastq
```

**5. Move the FASTQ files from your experiment into the newly created directory.**

```shell
# move all FASTQ files into it, e.g.
mv /path/to/experiment/*fastq.gz fastq/.
```
Next, the `config.yaml` file needs to be modified with the project-specific parameters and sample list. The pipeline is now ready to start. Before running, we can perform a “dry” run which prints the entire execution plan.

6. Preview the pipeline execution with the `-n` and `-p` flags.

```bash
# shell
snakemake -np
```

7. Run the pipeline, utilizing multiple cores with the `--cores` flag.

```bash
# shell
snakemake -p --cores 8
```

4.2.5 Structure of working directory

Here is what the home directory should look like after running the pipeline.

```
|--[Miniconda3]
  |--This is the working environment with all the tools installed inside.
  |--[genome]
      |--[gtf]
          |--gcencode.v25.annotation.gtf
      |--[fasta]
          |--GRCh38.p7.genome.fa
      |--[index]
          |--[rsem]
              |--This folder contains the RSEM genome reference files.
  |--[rnaseq_workflow]
      |--requirements.yaml
      |Snakefile
      |--config.yaml
    |--[fastq]
      |--sample1_1.fastq.gz
      |--sample1_2.fastq.gz
      |--sample2_1.fastq.gz
      |--sample2_2.fastq.gz
      ...
    |--[fastQC]
      |--[sample1]
          |--stdin_fastqc.html
          |--stdin_fastqc.zip
      |--[samples2]
          |--stdin_fastqc.html
```
During RNA-Seq data processing, it is difficult to keep track of all the input, intermediate and output files. The process is also likely to take a long time because of the amount of the data. With this automated pipeline, the inputs and outputs can be placed in the project folder in a highly organized manner. The pipeline can also be suspended and resumed at any time without having to re-run it from the beginning. For each project, we only need to modify the parameters in the configuration file, which is much easier than rewriting scripts every time we start a new project, thereby greatly speeding up our workflow. In addition, since the working instructions, software (including version numbers), and software parameters are stored automatically, the pipeline also ensures the reproducibility of the data if we ever want to return to a project.
Reference


Appendices

Appendix A  Supplementary material for Chapter 2

A.1  Genes of interest (microarrays) comparing ER and DR at pre-challenge (8ERs, 6DRs)

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Gene Symbol</th>
<th>logFC*</th>
<th>Average Expression</th>
<th>t</th>
<th>P-Value</th>
<th>BH-FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>8037374</td>
<td>PLAUR</td>
<td>-0.350</td>
<td>8.040</td>
<td>-2.360</td>
<td>0.031</td>
<td>0.854</td>
</tr>
<tr>
<td>8029907</td>
<td>C5AR1</td>
<td>-0.286</td>
<td>9.032</td>
<td>-1.780</td>
<td>0.094</td>
<td>0.854</td>
</tr>
<tr>
<td>8003656</td>
<td>SERPINF2</td>
<td>0.218</td>
<td>5.032</td>
<td>1.770</td>
<td>0.095</td>
<td>0.854</td>
</tr>
<tr>
<td>7981068</td>
<td>SERPINA1</td>
<td>-0.295</td>
<td>9.906</td>
<td>-1.743</td>
<td>0.100</td>
<td>0.854</td>
</tr>
<tr>
<td>8123259</td>
<td>PLG</td>
<td>0.425</td>
<td>4.335</td>
<td>1.704</td>
<td>0.107</td>
<td>0.854</td>
</tr>
<tr>
<td>8170215</td>
<td>F9</td>
<td>0.163</td>
<td>2.228</td>
<td>1.644</td>
<td>0.119</td>
<td>0.854</td>
</tr>
<tr>
<td>8024062</td>
<td>CFD</td>
<td>-0.437</td>
<td>9.322</td>
<td>-1.548</td>
<td>0.141</td>
<td>0.854</td>
</tr>
<tr>
<td>7909400</td>
<td>CD46</td>
<td>-0.229</td>
<td>8.986</td>
<td>-1.545</td>
<td>0.141</td>
<td>0.854</td>
</tr>
<tr>
<td>8159491</td>
<td>C8G</td>
<td>-0.197</td>
<td>6.013</td>
<td>-1.545</td>
<td>0.142</td>
<td>0.854</td>
</tr>
<tr>
<td>7947425</td>
<td>CD59</td>
<td>-0.291</td>
<td>6.752</td>
<td>-1.535</td>
<td>0.144</td>
<td>0.854</td>
</tr>
</tbody>
</table>

*: mean log$_2$(expression in DR) - mean log$_2$(expression in ER)
A.2 Proteins of interest (iTRAQ proteomics) comparing ER and DR at pre-challenge (4ERs, 4DRs)

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Protein Name</th>
<th>FC*</th>
<th>p-Value</th>
<th>BH-FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1R</td>
<td>cDNA FLJ54471, highly similar to Complement C1r subcomponent</td>
<td>0.75</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>C1QC</td>
<td>Complement C1q subcomponent subunit C</td>
<td>0.63</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>C1S</td>
<td>Complement C1s subcomponent</td>
<td>0.67</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>C7</td>
<td>Complement component C7</td>
<td>0.83</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>C8G</td>
<td>Complement component C8 gamma chain</td>
<td>0.87</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>C1QB</td>
<td>Complement component 1, q subcomponent, B chain</td>
<td>0.82</td>
<td>0.04</td>
<td>0.16</td>
</tr>
<tr>
<td>C5</td>
<td>Complement component C5</td>
<td>0.61</td>
<td>0.05</td>
<td>0.16</td>
</tr>
</tbody>
</table>

* Fold-change: average expression in DR / average expression in ER

A.3 NanoString quality control (QC) criteria

1. Imaging QC: % FOV (field of view) must be greater than 75 FOV.
2. Binding Density QC: Binding density must be between 0.05 and 2.25.
3. Positive Control Linearity QC: The R² must be greater than 0.9 between the counts and concentrations of the 6 positive controls.
4. Positive Control Limit of Detection QC: The second lowest positive control spike in (0.5fM) must have counts greater than the Mean ± 2SD of the negative controls for each sample.
## A.4 NanoString probe sequence of candidate genes

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Accession</th>
<th>Position</th>
<th>Target Sequence</th>
<th>Tm CP</th>
<th>Tm RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3AR1</td>
<td>NM_0040</td>
<td>416-515</td>
<td>CAGTGTCTTCTCTGCTTA...</td>
<td>80</td>
<td>81</td>
</tr>
<tr>
<td>C5AR1</td>
<td>NM_0017</td>
<td>1196-1295</td>
<td>TTGCCTGTCTTTCCAGACTTGTCC...</td>
<td>79</td>
<td>82</td>
</tr>
<tr>
<td>CD46</td>
<td>NM_1723</td>
<td>366-465</td>
<td>TATACCTCTCTTCCGCCACCCCATACTATTTG...</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>CD55</td>
<td>NM_0005</td>
<td>102-201</td>
<td>CCCTACTCCACCCGTCTTTGTTCCT...</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>CD59</td>
<td>NM_0006</td>
<td>731-830</td>
<td>GACTTGAACTAGATTGCATGCTTCTCT...</td>
<td>81</td>
<td>78</td>
</tr>
<tr>
<td>F13A1</td>
<td>NM_0001</td>
<td>3197-3296</td>
<td>TTCAGGTCCCTTTCCAGAGATATAAT...</td>
<td>80</td>
<td>85</td>
</tr>
<tr>
<td>PLAUR</td>
<td>NM_0010</td>
<td>441-540</td>
<td>GAGAAGACCAACAGGACCCTCTAGCTGCT...</td>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>SERPINA1</td>
<td>NM_0002</td>
<td>761-860</td>
<td>TCACTGTCAAATCTCCGGGAGACACCAG...</td>
<td>81</td>
<td>79</td>
</tr>
</tbody>
</table>

Tm CP: melting temperature of the capture probe

Tm RP: melting temperature of the reporter probe
Appendix B  Supplementary material for Chapter 3

B.1  Discovery and validation workflow of the WRCA panel
B.2 Classification tree of the best performing panel
### List of significant genes comparing expression change from 0h to 6h (PA-positive vs. PA-negative)

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>t</th>
<th>P-Value</th>
<th>BH-FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATA3</td>
<td>-4.090</td>
<td>0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>ENTPD5</td>
<td>4.030</td>
<td>0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>CHP1</td>
<td>3.996</td>
<td>0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>CARM1</td>
<td>3.800</td>
<td>0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>HIP1</td>
<td>3.740</td>
<td>0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>SPARC</td>
<td>3.662</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>SULT1A2</td>
<td>3.644</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>PELI1_isoform</td>
<td>3.618</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>NAPA</td>
<td>3.573</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>IL17A</td>
<td>3.530</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>unknown2a_comp55647_c0_seq2</td>
<td>-3.522</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>SH3BGRL3</td>
<td>3.493</td>
<td>0.002</td>
<td>0.030</td>
</tr>
<tr>
<td>MPPED1</td>
<td>3.484</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>PSMF1</td>
<td>3.445</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>RALGPS2_isoform</td>
<td>3.423</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>C9orf78</td>
<td>3.399</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>HBA2</td>
<td>3.384</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>PDCD1</td>
<td>3.362</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>F13A1</td>
<td>3.342</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>MBNL3</td>
<td>3.277</td>
<td>0.004</td>
<td>0.034</td>
</tr>
<tr>
<td>RRAD</td>
<td>3.236</td>
<td>0.004</td>
<td>0.035</td>
</tr>
<tr>
<td>GTF2H2</td>
<td>3.018</td>
<td>0.007</td>
<td>0.054</td>
</tr>
<tr>
<td>MAPKAPK2</td>
<td>2.903</td>
<td>0.009</td>
<td>0.065</td>
</tr>
<tr>
<td>unknown2_comp55647_c0_seq2</td>
<td>2.888</td>
<td>0.010</td>
<td>0.065</td>
</tr>
<tr>
<td>noName</td>
<td>ncRNA</td>
<td>2.879</td>
<td>0.010</td>
</tr>
<tr>
<td>RPS6</td>
<td>-2.845</td>
<td>0.010</td>
<td>0.066</td>
</tr>
<tr>
<td>MT-ND1</td>
<td>2.834</td>
<td>0.011</td>
<td>0.066</td>
</tr>
<tr>
<td>AKT3</td>
<td>-2.777</td>
<td>0.012</td>
<td>0.072</td>
</tr>
<tr>
<td>DAP</td>
<td>2.760</td>
<td>0.013</td>
<td>0.072</td>
</tr>
<tr>
<td>SF3B1</td>
<td>-2.720</td>
<td>0.014</td>
<td>0.076</td>
</tr>
<tr>
<td>MT-CYB</td>
<td>2.672</td>
<td>0.015</td>
<td>0.081</td>
</tr>
<tr>
<td>ATP8A1</td>
<td>-2.650</td>
<td>0.016</td>
<td>0.083</td>
</tr>
<tr>
<td>RALGPS2</td>
<td>-2.592</td>
<td>0.018</td>
<td>0.087</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>t</td>
<td>P-Value</td>
<td>BH-FDR</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>SCARNA5</td>
<td>-2.575</td>
<td>0.019</td>
<td>0.087</td>
</tr>
<tr>
<td>FOS</td>
<td>2.573</td>
<td>0.019</td>
<td>0.087</td>
</tr>
<tr>
<td>PPP3R1</td>
<td>2.572</td>
<td>0.019</td>
<td>0.087</td>
</tr>
</tbody>
</table>
B.4 Expression levels of the significant genes comparing expression change from 0h to 6h (PA-positive vs. PA-negative)