Network and machine learning approaches to dengue omics data

Iryna Nikolayeva

University Paris-Descartes, Doctoral School FDV (ED474)

Systems Biology Lab and Functional Genetics of Infectious Diseases Unit, Pasteur Institute

Ph.D. thesis

Network and machine learning approaches to dengue omics data

Iryna NIKOLAYEVA

Supervisors: Benno SCHWIKOWSKI and Anavaj SAKUNTABHAI

Reviewers :	Jean-Philippe VERT	-	PSL University, MINES ParisTech, ENS Paris, Institut Curie, INSERM
	Zohar YAKHINI	-	Herzeliya Interdisciplinary Center, Technion
Examinators :	Kristel VAN STEEN	-	Universities of Liege (ULg), K.U Leuven, Ghent
	Etienne PATIN	-	Institut Pasteur, CNRS
	Marie FLAMAND	-	Institut Pasteur

Abstract

The last 20 years have seen the emergence of powerful measurement technologies, enabling omics analysis of diverse diseases. They often provide non-invasive means to study the etiology of newly emerging complex diseases, such as the mosquito-borne infectious dengue disease. My dissertation concentrates on adapting and applying network and machine learning approaches to genomic and transcriptomic data.

The first part goes beyond a previously published genome-wide analysis of 4,026 individuals by applying network analysis to find groups of interacting genes in a gene functional interaction network that, taken together, are associated to severe dengue. In this part, I first recalculated association p-values of sequences polymorphisms, then worked on mapping polymorphisms to functionally related genes, and finally explored different pathway and gene interaction databases to find groups of genes together associated to severe dengue.

The second part of my dissertation unveils a theoretical approach to study a size bias of active network search algorithms. My theoretical analysis suggests that the best score of subnetworks of a given size should be size-normalized, based on the hypothesis that it is a sample of an extreme value distribution, and not a sample of the normal distribution, as usually assumed in the literature. I then suggest a theoretical solution to this bias.

The third part introduces a new subnetwork search tool that I co-designed. Its underlying model and the corresponding efficient algorithm avoids size bias found in existing methods, and generates easily comprehensible results. I present an application to transcriptomic dengue data.

In the fourth and last part, I describe the identification of a biomarker that detects dengue severity outcome upon arrival at the hospital using a novel machine learning approach. This approach combines two-dimensional monotonic regression with feature selection. The underlying model goes beyond the commonly used linear approaches, while allowing to control the number of transcripts in the biomarker. The small number of transcripts along with its visual representation maximize the understanding and the interpretability of the biomarker by biomedical professionals. I present an 18-gene biomarker that allows distinguishing severe dengue patients from non-severe ones upon arrival at the hospital with a unique biomarker of high and robust predictive performance. The predictive performance of the biomarker has been confirmed on two datasets that both used different transcriptomic technologies and different blood cell subtypes.

A mes parents,

A mes grands-parents,

... And to all those who dare to (re)search.

Contents

Acknowledgments

"50% of the success of your PhD is your work on your subject, and the other 50% is your relationship with your supervisor and people in your environment". This was the best advice that I had received before starting my PhD, and that appeared to be so true. I would like to thank my supervisors Dr. Benno Schwikowski and Dr. Anavaj Sakuntabhai for their effort to create and navigate through an interdisciplinary collaboration. Thank you Anavaj for providing diverse questions, data, for being open to suggestions, and for your patience when the results take time to appear. Thank you for your trust in the biomarker results and your willingness to validate them. Thank you Benno for your everyday support during different technical issues and philosophical doubts, thank you for your openness, your adaptivity, and your kind comments that helped me build trust in my capabilities. Thank you for your time during pre-deadline evenings and weekends. Thank you to the thesis advisory committee members, Dr. Etienne Patin and Dr. Kristel van Steen, for taking the time to read, listen, and give very useful feedback during yearly progress meetings. They helped me greatly to prioritise projects through the PhD. Kristel, thank you for the research on eQTLs that was done in your lab, as well as an acquired taste for statistics, and thanks to your team members for hosting me in you lab and allowing me to discover a new environment.

I thank the financial support of LabEx IBEID, the European project HERCULES, Institut OpenHealth, and the doctoral school FDV, as well as the administrative support of the personnel of Institut Pasteur.

Thank you to past and present members of the Systems Biology team Robin Friedman and Frederik Gwinner for your clever advices, Freddy Cliquet for your amazing computer and cluster fixing magic, Oriol Guitart-Pla for your support with computer programming, and last but not least Pierre Bost who transformed not only the interpretation of my results thanks to a broad immunological knowledge, but also gave me a new taste for this job. Thank you the members of the GFMI lab and more specifically Isabelle Casademont, Ahmed Tawfik Moustafa, Etienne Simon-Lorière, Jean-François Bureau, for always being welcoming and for your help. A special thank you to Laura Grange for providing a very

enjoyable first collaboration and good advice from PhD student to PhD student. Thanks Kevin for your critical eye on the statistical methods that prevented us from doing several mistakes!

I would like to thank my doctoral school, that in my opinion followed greatly Plutarch's assertion that "The mind is not a vessel to be filled, but a fire to be kindled" by allowing me to explore also those other subjects that interest me whenever this interest appears. Thank you for your support for all the students' multidisciplinary interests, and for allowing me to develop my interest for education. I need to continue by thanking the sAvanturiers team, Fanny Peissik, Justine Hannequin, Marine Lehue for contributing in two exciting science projects with CE1 and CE2 classes. These moments give so much energy, smiles, and connection! I would like to also thank François Taddei for following the students of the doctoral school despite your big responsibilities. Thank you for your availability and advice! Thank you Allon and the playbackers for your craziness, openness, willingness to share, create and for your warmth!

... Et tous ceux auprès desquels j'ai fait ce long cheminement. Merci en premier lieu à mes parents: se savoir aimée, écoutée et pouvoir compter sur vous est le plus beau des cadeaux. Merci de m'avoir permis d'arriver jusque-là. Maks, grandir ensemble est un grand plaisir! Merci aux mamies Nathalie et Margaryta pour votre soutien infaillible de près ou de loin! Merci Aurélien pour ta façon de m'épauler lors de cette looongue écriture contre vents et marées... parfois non prévues par la météo thèse. Merci Aymeric pour ton aide statistique inestimable et ta présence aux instants critiques, festifs et lors des repas des mercredis soirs régulièrement interminables. Et aux chères colooooocs Juliette, Anouk et Irène de Rennes pour les moments mangeants, dansants, des baby-pauses ou de la redéfinition du monde ou des table basses! Merci à la bande silent disco et celle de Picherande. Ces projets sont magiques et redonnent force et courage! Un merci tout particulier à la team Nowhere et au barrio Espace sans l'électricité et la compréhension desquels cette thèse ne serait pas allée aussi loin à la découverte de la planète immunologie. Merci au trio FDV représente pour le soutien "thésards pas si anonymes!" parce que tous dans la même galère et les mêmes festoiements! Et merci Tristan Cerisy pour ton formidable conseil des 50%.

Chapter 1

Introduction

Contents

This chapter introduces notions that are required to understand how I link measurements of molecular features to dengue disease. I start by describing the disease itself. Then, I will very briefly describe the types of data that I am analysing. Finally, I will introduce the basics of the computational methods I employed for data analysis: interaction network-based and machine learning algorithms.

1.1 Dengue, a complex disease

1.1.1 Epidemiology

Dengue is the most widespread mosquito-borne viral infection worldwide. Currently, an estimated 40% to 50% of the world population lives in areas where the mosquito transmitting the virus has spread, and are therefore at risk for dengue virus transmission [WHO, 2017]. Figure 1.1 shows countries that are now considered to be at risk for a dengue epidemic. The dengue virus is closely related to the Zika virus in terms of symptoms of infection, transmission and even protein structure [Priyamvada et al., 2016]. The recent increase in spread and virulence of Zika gives an example of potential dangers that dengue may represent in the close future.

Dengue was first recognized in the 1950s during epidemics in the Philippines and Thailand. Since then, its incidence has grown fast. Before 1970, only nine countries had experienced severe dengue epidemics. The disease is now endemic in more than 60 countries in Africa, the Americas, the Eastern Mediterranean, South-east Asia and the Western Pacific. The American, South-east Asia and the Western Pacific regions are the most seriously affected. Recently the number of reported cases has continued to increase. An estimated 500,000 people with severe dengue require hospitalisation each year, with a large proportion of severe cases occurring in children and the elderly. About 2.5% of those affected die [WHO, 2017]. Not only is the number of cases increasing as the disease spreads to new areas, but explosive outbreaks occur. The threat of a possible outbreak now exists in Europe, and local transmission of dengue was reported for the first time in France and Croatia in

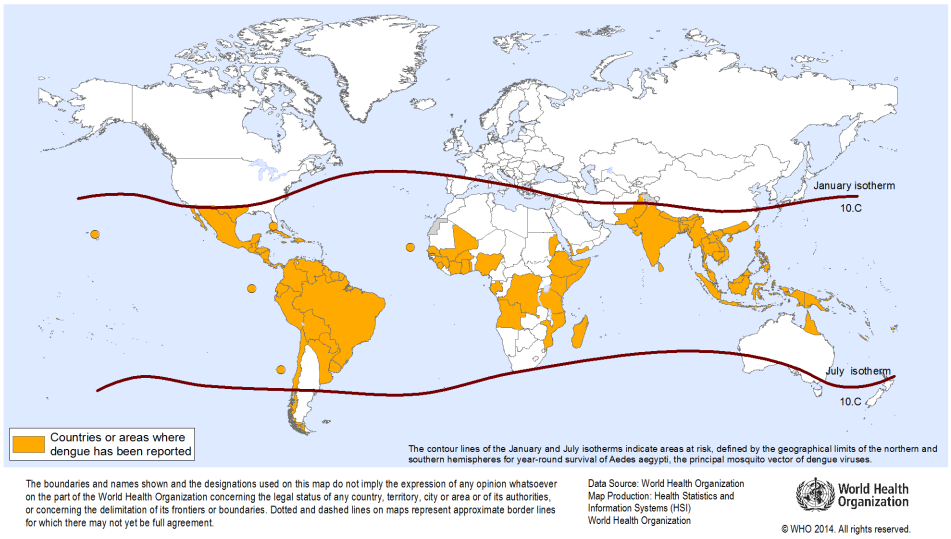


Figure 1.1: Dengue: countries or areas at risk of a dengue epidemic based on the most recent consensus. The countries in orange had dengue epidemics reported before year 2013, while the countries in between the two isotherms have a climate adapted to the main mosquito vector that transmits dengue. Source: WHO, 2014.

2010. Imported cases are regularly detected during holiday periods in European countries, including France.

1.1.2 Transmission

The *Aedes aegypti* mosquito is the primary vector of dengue. The virus is transmitted to humans through the bites of infected female mosquitoes. After virus incubation for 4–10 days, an infected mosquito is capable of transmitting the virus for the rest of its life. Infected humans are the main carriers and multipliers of the virus, serving as a source of the virus for uninfected mosquitoes. Patients who are already infected with the dengue virus can transmit the infection (for 4–5, maximally 12 days) via *Aedes* mosquitoes after their first symptoms appear. The *Aedes aegypti* mosquito lives in urban habitats and breeds mostly in man-made containers. It is thus very adapted to big concentrations of human population, such as cities. *Aedes albopictus*, a secondary dengue vector in Asia, has spread to North America and Europe, largely due to the international trade in used tyres (a breeding habitat)

and other goods (i.e., lucky bamboo). *Aedes albopictus* is highly adaptable and can survive in the cooler temperate regions of Europe. Its spread is due to its tolerance to temperatures below freezing, hibernation, and its ability to find shelter in microhabitats.

1.1.3 Symptoms and severity classification

Reactions to infection by dengue virus can vary a lot from no symptoms, over flu-like symptoms, to deadly complications. Dengue is suspected when a high fever (40°C/104°F) is accompanied by two of the following symptoms: severe headache, pain behind the eyes, muscle and joint pains, nausea, vomiting, swollen glands, or rash. Due to the lacking specificity of some of these symptoms, dengue needs to be confirmed in the laboratory for a precise diagnostic. Symptoms usually last for 2–7 days, after an incubation period of 4–10 days after the bite from an infected mosquito. Severe dengue is a potentially deadly complication due to plasma, which leaks out of the vessels and into the organs, provoking fluid accumulation in the body cavities, respiratory distress, severe bleeding (because of the lack of platelets in which leak out with the plasma), potential organ impairment such as problems with liver or the nervous system, and, eventually, shock (i.e., a state where the heart ceases to correctly function, and stops). Warning signs occur 3–7 days after the first symptoms in conjunction with a decrease in temperature (below 38°C) and include: severe abdominal pain, persistent vomiting, rapid breathing, bleeding gums, fatigue, restlessness, blood in vomit. The next 24–48 hours of the critical stage can be lethal; proper medical care is needed to avoid complications and the risk of death.

1.1.4 Dengue severity classifications

Reactions to infection by dengue virus have a wide range of clinical manifestations and severities, from no symptoms to deadly complications. The evolution of the disease over time is often very difficult to predict for clinicians. Severe disease is difficult to define, but this is an important concern since appropriate treatment may prevent patients from developing more severe clinical conditions [WHO (World Health Organisation), 2009]. To

help physicians distinguish between the different forms of dengue, a WHO committee developed guidelines for case classification in 1974. Based on studies of disease patterns in children in Thailand in the 1960s, these guidelines were then modified and re-issued several times [Hadinegoro, 2012], notably in 1997. Many reports state difficulties in the use of this classification, such as lacking suitability to regions outside of Asia. They were summarized in a systematic literature review [Bandyopadhyay et al., 2006]. The classification of dengue cases was subsequently revised by distinguishing between dengue with and without warning signs and severe dengue, as published in 2009 [WHO (World Health Organisation), 2009]. I worked with the 1997 and 2009 classifications, and present them in more detail.

The 1997 guidelines classified dengue into DF, DHF (Grades 1 and 2) and DSS (DHF Grades 3 and 4; Figures 1.2 and 1.3). The case diagnosis emphasised the need for laboratory confirmation. Studies have demonstrated an overlap between the case definitions of DF, DHF and DSS, supporting the concept of dengue as a continuous spectrum of disease, rather than distinct subforms [Deen et al., 2006, Phuong et al., 2004].

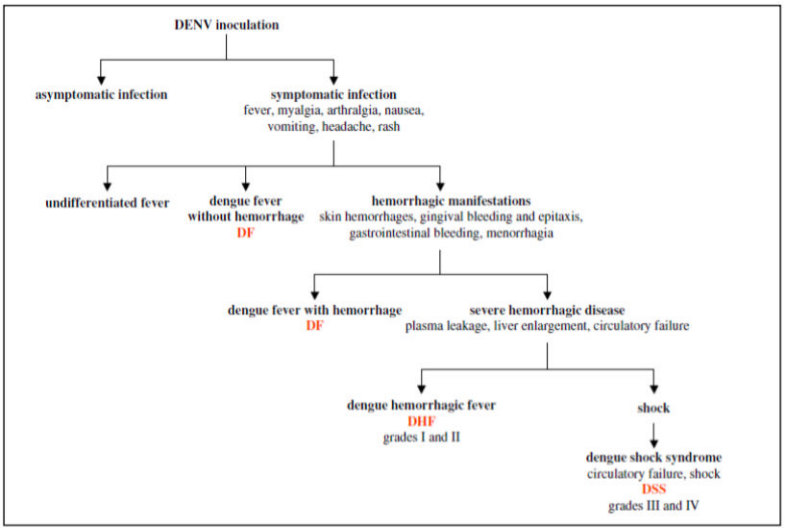


Figure 1.2: WHO 1997 classification. Source: [Grange, 2014]

The 2009 WHO criteria (Figure 1.4) classify dengue according to the following levels of severity: dengue without warning signs, dengue with warning signs (abdominal pain, persistent vomiting, fluid accumulation, mucosal bleeding, lethargy, liver enlargement, increas-

Grading severity of dengue haemorrhagic fever

DHF is classified into four grades of severity, where grades III and IV are considered to be DSS. The presence of thrombocytopenia with concurrent haemoconcentration differentiates grades I and II DHF from DF.

Grade I: Fever accompanied by non-specific constitutional symptoms; the only haemorrhagic manifestation is a positive tourniquet test and/or easy bruising.

Grade II: Spontaneous bleeding in addition to the manifestations of Grade I patients, usually in the forms of skin or other haemorrhages.

Grade III: Circulatory failure manifested by a rapid, weak pulse and narrowing of pulse pressure or hypotension, with the presence of cold, clammy skin and restlessness.

Grade IV: Profound shock with undetectable blood pressure or pulse.

Grading the severity of the disease at the time of discharge has been found clinically and epidemiologically useful in DHF epidemics in children in the WHO Regions of the Americas, South-East Asia and the Western Pacific, and experience in Cuba, Puerto Rico and Venezuela suggests that grading is also useful for adult cases.

Figure 1.3: WHO 1997 classification description

ing hematocrit with decreasing platelets), and severe dengue (dengue with severe plasma leakage, severe bleeding, or organ failure) [WHO (World Health Organisation), 2009]. The 2009 classification according to levels of severity has been considered to be more sensitive in capturing severe disease than the 1997 guidelines, with observed sensitivities of up to 92% for the 2009 WHO classification, against 39% for the 1997 WHO classification [Hadinegoro, 2012, Basuki et al., 2010, Narvaez et al., 2011].

1.1.5 Treatment

There is no anti-viral drug treatment for dengue fever. For severe dengue, medical care by physicians and nurses experienced with the effects and progression of the disease can save lives, decreasing mortality rates from more than 20% to less than 1%. Maintenance of body fluid volume via intravenous rehydration is critical to severe dengue care.

Many vaccine trials are currently being conducted [Dengue Vaccine Initiative, 2017]. One vaccine has recently passed clinical trials [Hadinegoro et al., 2015], is approved by 11 countries, but its efficacy is limited. Moreover, the scientific community wonders whether there

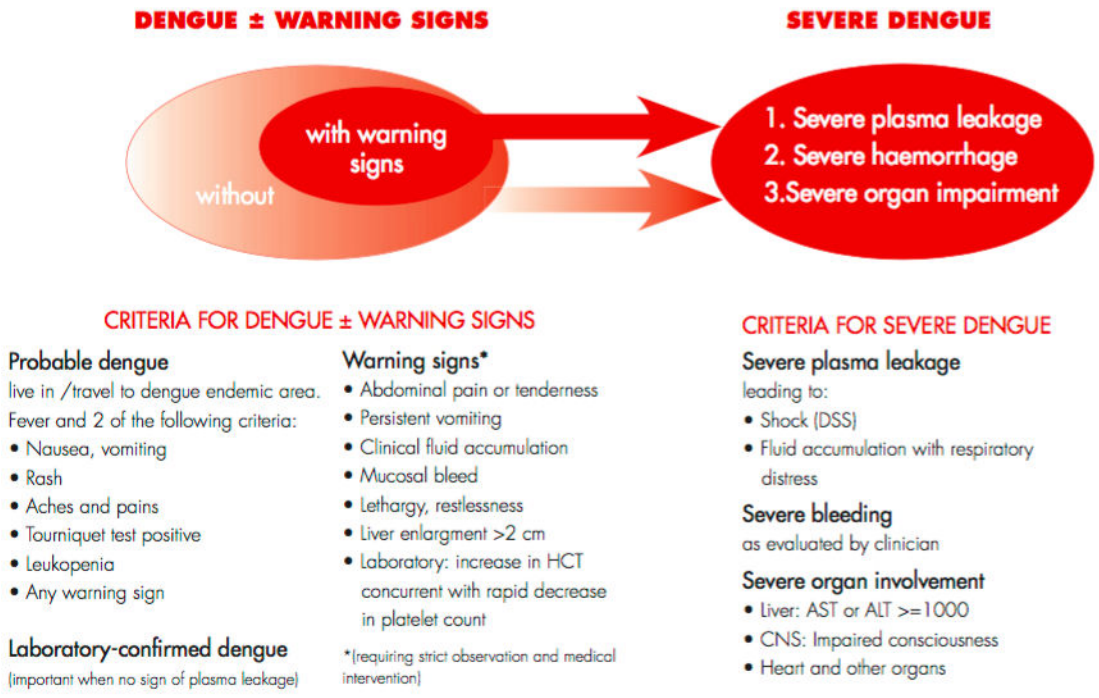


Figure 1.4: WHO 2009 classification

is a correlation between this vaccine and an increased probability of contracting severe Zika [Priyamvada et al., 2016].

At present, the main approach to control the transmission of dengue virus is to combat vector mosquitos, but sufficient mosquito control remains a challenge, and the disease is spreading quickly. This motivates the search of possible treatments using all contemporary tools.

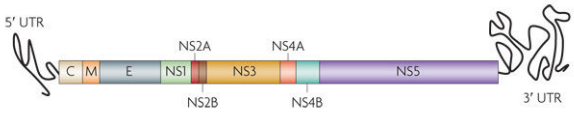
1.1.6 Dengue virology and immunopathology

Insights into the pathogenesis of severe dengue are hampered by the lack of an animal model that accurately recreates the transient capillary permeability syndrome, accompanied by a decreasing viral burden that is seen in severe patients [Simmons et al., 2012]. Therefore many discoveries remain to be validated and are frequently debated. This section gives a broad introduction of the current understanding of immunological processes implicated in

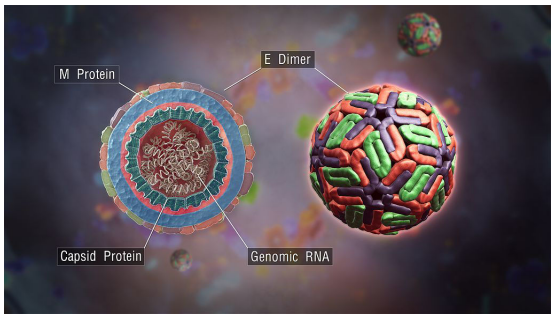
severe dengue.

The virus

Dengue is a single positive-stranded RNA of the family Flaviviridae, genus Flavivirus. Other members of this genus include Zika, yellow fever and West Nile virus. Dengue has four serotypes that have evolved in parallel in different places worldwide, and only recently co-exist in endemic countries. A fifth serotype has been reported in 2013, but has not yet been confirmed by independent studies [Normile, 2013]. Figure 1.5 presents the proteins of the virus.



(a) Proteins of the dengue virus polyprotein
Source: [Guzman et al., 2010]



(b) Mature dengue virion.
Source: http://www.scientificanimations.com/

Figure 1.5: Proteins of the dengue virus polyprotein. The single positive-stranded RNA, codes for three structural proteins (capsid protein C, membrane pre-M protein that will mature into an M protein when travelling in the virion, envelope protein E) and seven nonstructural proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b, NS5). NS1 has a specific role in the modulation of the immune reaction, as will be explained later on. Dengue RNA also includes short non-coding regions on both the 5' and 3' ends. [Rothman, 2011].

Where and how is the RNA translated into this polyprotein? Dengue virions bind to cell surface receptors of immune cells, such as monocytes, macrophages or dendritic cells, and are internalised through endocytosis. Acidification of the endocytic vesicle leads to rearrangement of the surface envelope (E) glycoprotein, fusion of the viral and vesicle membranes, and release of viral RNA into the cytoplasm. Host translational machinery (ribosomes) translates the RNA into a single polypeptide. Cellular and viral proteinases cleave the

polypeptide into 10 proteins (E, M, C and 7 non-structural/enzymatic proteins). The viral proteins and newly synthesized viral RNA assemble into immature virions within the ER lumen. As soon as functional RNA-dependent RNA polymerase is synthesised, RNA replication can start. A negative strand of the RNA is generated from the positive one. From this negative strand intermediate, a new positive strand is generated. This process generates 10 times more copies of the positive strand than of the negative.

Cleavage of the viral precursor membrane (pre-M) protein by the host cell enzyme furin leads to the formation of mature virions, which are secreted from the cell. In addition, some of the synthesized non-structural protein 1 (NS1) is expressed on the plasma membrane of the cell or secreted, and some virions are secreted in an immature form.

Immunopathology of dengue disease

When an infected mosquito feeds on a person, it injects the dengue virus into the bloodstream. The virus infects nearby skin cells called keratinocytes, the most common cell type in the skin. The dengue virus also infects and replicates inside a specialized immune cell located in the skin, a type of dendritic cell called a Langerhans cell. The virus enters the cells by binding to membrane proteins on the Langerhans cell, specifically DC-SIGN, mannose receptor and CLEC5A [Rodenhuis-Zybert et al., 2010]. DC-SIGN, a non-specific receptor for foreign material on dendritic cells, seems to be the main point of entry [Guzman et al., 2010]. The Langerhans cells then maturate, travel to the lymph nodes and alert the immune system to trigger the immune response because a pathogen is in the body. In the meantime, the virus replicates in the Langerhans cells and is released into the bloodstream. Once in the bloodstream, it can infect several other blood leukocytes such as monocytes and macrophages.

When the virus infects immune cells, it uses its machinery to replicate and be released from these cells, while the cells emit inflammatory signals such as cytokines (including interferons type I and II) to trigger the immune defense reaction. This inflammation becomes systemic when the virus spreads in the body and causes most of the severe dengue symptoms. Figure

1.6 illustrates the time evolution of severe dengue as well as causes and consequences of systemic inflammation in diseased secondary dengue patients. The inflammation triggers the reaction of the immune system via T-cells, the complement system, and antibodies simultaneously. We will further explain each of these immune reactions and their consequences on the pathology. We will then present a very specific property to secondary infection by dengue, called antibody dependent enhancement (ADE).

T-cell response

As previously indicated, the infected macrophage or dendritic cell is an antigen presenting cell (APC). It presents antigens on its surface via the MHC class I and II molecules. Cytotoxic T-cells, also known as $CD8^+$, bind to MHC class I and lyse the infected cell. T helper cells, also known as $CD4^+$, bind to MHC class II, release additional inflammatory cytokines and assist other immunologic processes, including maturation of B cells into plasma cells. This maturation enables them to produce many neutralizing antibodies, trigger the antibody response, and activate cytotoxic T cells and macrophages to lyse the infected cells (Figure 1.7).

This system becomes less efficient if the presented antigen resembles one that had already been encountered, but has a slightly modified shape. This is the case for a secondary infection with a new dengue virus serotype, and is known as the "original antigenic sin" [Francis, 1960].

The complement

The complement is a complex system of more than 30 proteins that are part of the innate immune response. The interacting proteins of the complement system, which are produced mainly by the liver, circulate in the blood and extracellular fluid, primarily in an inactivated state. Not until the system receives an appropriate signal are they activated. The signal sets off a chemical chain reaction in which cleaved complement proteins trigger the cleavage of the next complement protein in the sequence [Martina et al., 2009].

Complement activation occurs in dengue either by the *classical pathway* or the *alterna-*

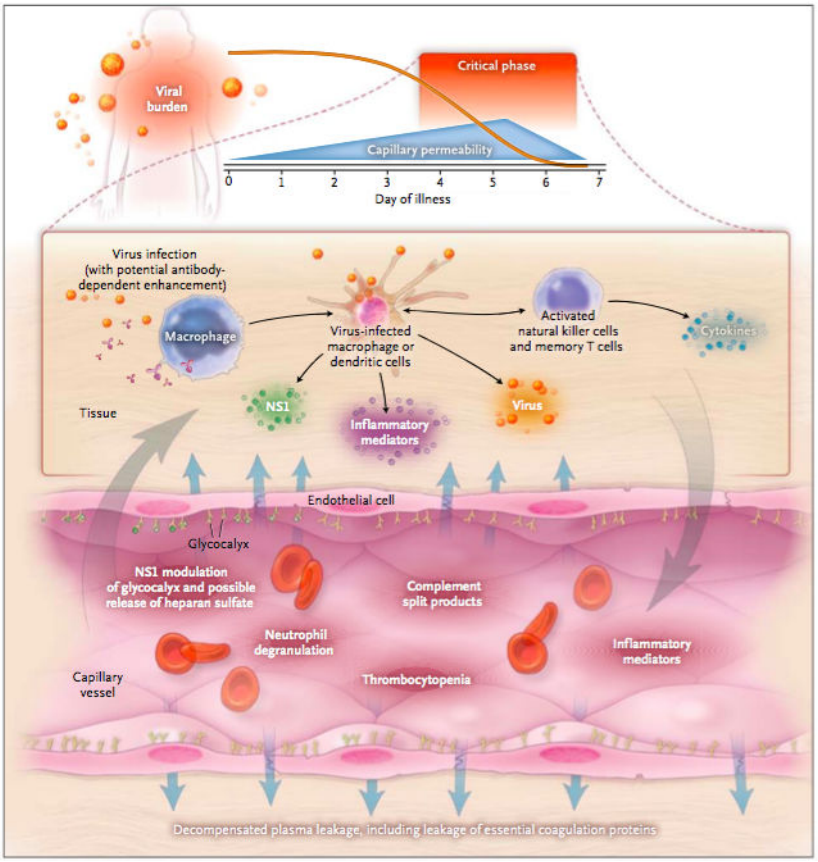


Figure 1.6: Immunopathogenesis of severe dengue in secondary patients. The kinetics of viral burden (i.e., concentration of the virus in blood), the timing of common complications, and possible mechanistic causes are shown. During the most severe, possibly life-threatening, critical phase, the viral burden decreases. The strong immune reaction is responsible of the most severe symptoms. A large infected cell mass results in elevated systemic concentrations of acute-phase response proteins, cytokines, and chemokines, and generation of antibody-antigen aggregates, immune complexes. Collectively, the host immunologic response is thought to create a physiological environment in tissues that promotes capillary permeability (via the interaction of a viral protein NS1, with the capillary epithelial glycocalix, which results in release of heparan sulfate, that, in turn, increases permeability). Loss of essential coagulation proteins such as platelets probably plays a major role in the development of bleeding-related symptoms. Source: [Simmons et al., 2012].

tive pathway. A different type of signal activates each pathway. The classical pathway is triggered by groups of antibodies bound to the surface of microorganisms. The alternative pathway is spurred into action by molecules embedded in the surface membranes of invading microorganisms, and does not require the presence of antibodies. Both pathways converge

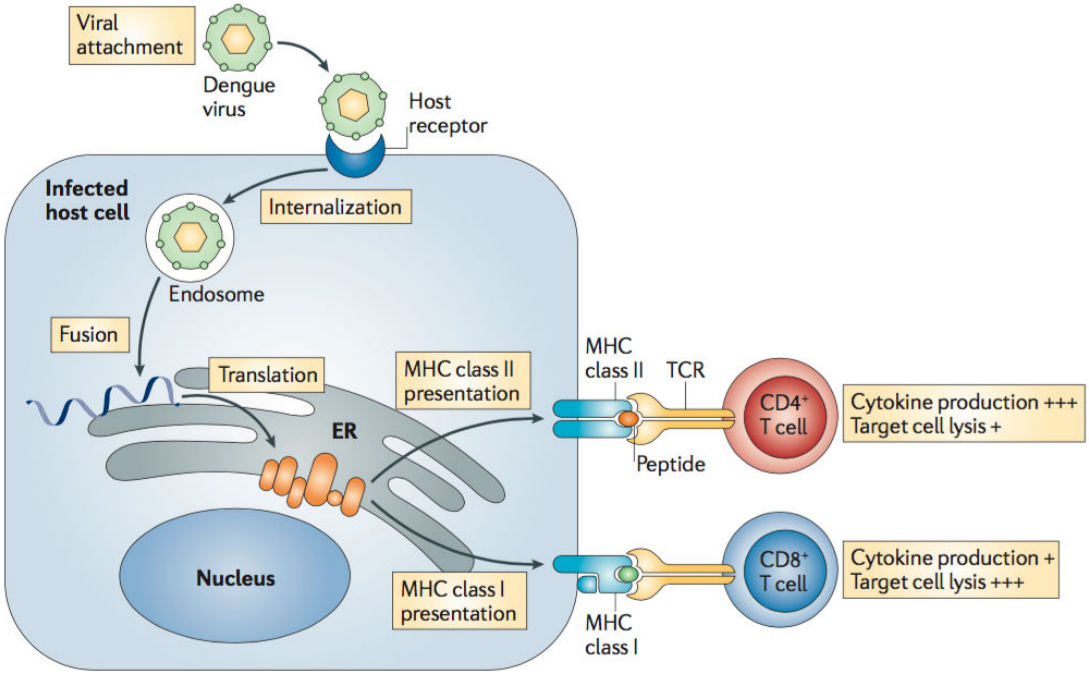


Figure 1.7: T-cell response to dengue infection. Source: [Rothman, 2011].

to activate the pivotal protein of the complement system, called C3. Once activated, the complement system causes lysis of infected cells, phagocytosis of foreign particles, as well as cell debris and the inflammation of surrounding tissue.

With regard to dengue, it was noticed that, around the time of defervescence in severe patients, when plasma leakage may become apparent, high levels of the activation products C3a and C5a are present in the plasma, followed by an accelerated consumption and a marked reduction of the complement components [Churdboonchart et al., 1983, Shaio et al., 1992]. Therefore, it was hypothesized that complement activation plays an important role in the pathogenesis of severe dengue. Comparison of global gene expression profiles in peripheral blood mononuclear cells of severe versus non-severe dengue patients also suggests the involvement of the complement system in disease severity [Ubol et al., 2008]. However, many aspects of complement activation and its role in dengue pathogenesis remain to be investigated. It has been proposed that binding of antibodies to NS1 expressed on infected cells may result in complement activation [Avirutnan et al., 2006, Lin et al., 2008] (Figure

1.8).

Antibody response

In parallel to the T-cell mediated immune response, B-cell mediated immunity is triggered during the course of dengue infection and results into the production of a large amount of virus-neutralizing IgG antibodies. In the case where the virus has not been previously encountered by the immune system, some naive B cells will be able to bind the virus through their B cell receptor (BCR; a membrane form of the antibody), and with the help of specific T cells, will differentiate into plasma cells inside the lymph node. During this differentiation, the affinity of the germline-encoded BCR will increase through the hypersomatic mutation process and B cells start to produce large amounts of IgG antibodies that will neutralise the virus.

In the case of secondary infection, memory B cells and persistent plasma cells will quickly produce large amount of IgG antibodies without the help of T cells.

When the matching IgG antibodies are released into blood, they specifically recognize and neutralize the dengue viral particles (Figure 1.8) as well as improve the efficiency of phagocytosis via their Fc region.

Antibody-dependent enhancement (ADE)

Once infected by one dengue serotype, the organism acquires a lifelong protection against any future infection by this serotype and a several weak immunity for all other serotypes. But a very remarkable, and to my knowledge unique, fact is that once the immunity against other serotypes is lost, the risk of developing severe dengue during secondary infection (i.e. when infected for a second time by an other dengue serotype) increases [Halstead, 2003]. There has been a lot of research done in order to understand why the secondary reaction is more severe. A detailed review was recently published [Screaton et al., 2015]. The most well-studied mechanism causing this severe reaction is known as antibody-dependent enhancement (ADE) [Sangkawibha et al., 1984].

As previously said, during a secondary infection by dengue, patients possess antibodies that are adapted to the previously encountered viral serotype.

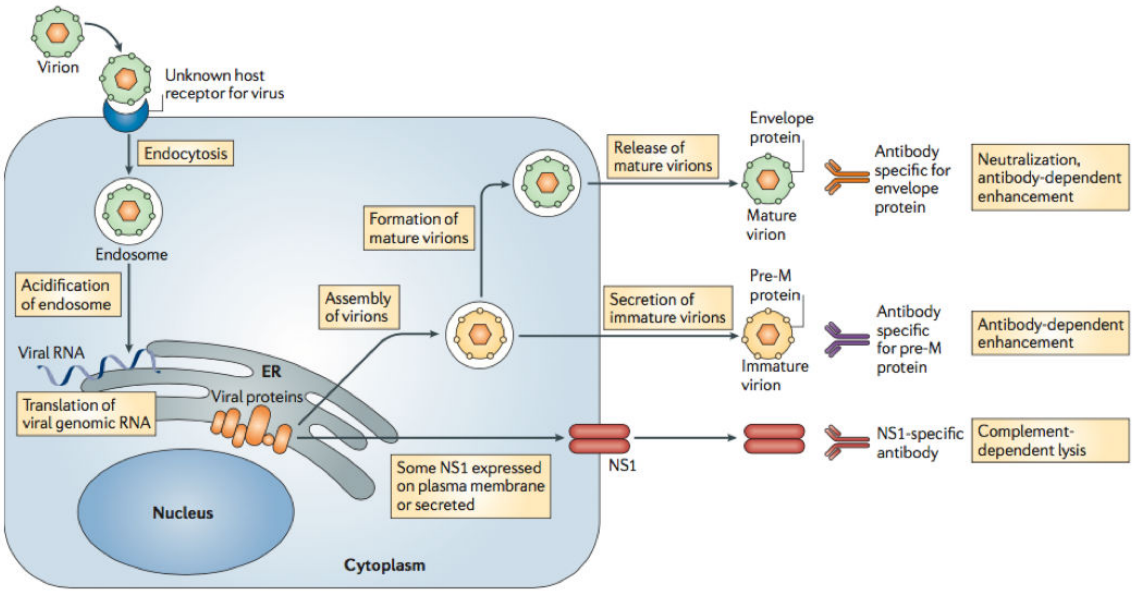


Figure 1.8: Dengue virus life cycle and antibody response to the pathogen. Mature and immature virions induce antibody responses to the E protein, and these antibodies can function in neutralisation or in antibody-dependent enhancement of infection. Immature virions also induce antibody responses to the pre-M protein. Antibodies specific for NS1 can interact with membrane-bound NS1 and cause complement-dependent lysis of virus-infected cells. Source: [Rothman, 2011].

Antibodies specific to the exact virus serotype completely block virion entry into the cell. Antibodies that do not match the exact serotype bind only incompletely; the virion is able to penetrate easily the phagocytic immune cell, thanks to the recognition of the Fc part of the antibody by the Fc gamma receptor, and the antibodies do not prevent it from replicating once in the immune cell. Therefore, if antibody binding is incomplete, the virus actually penetrates easier inside the host cell, and thus replicates more easily (Figure 1.9). This phenomenon is called antibody-dependent enhancement (ADE).

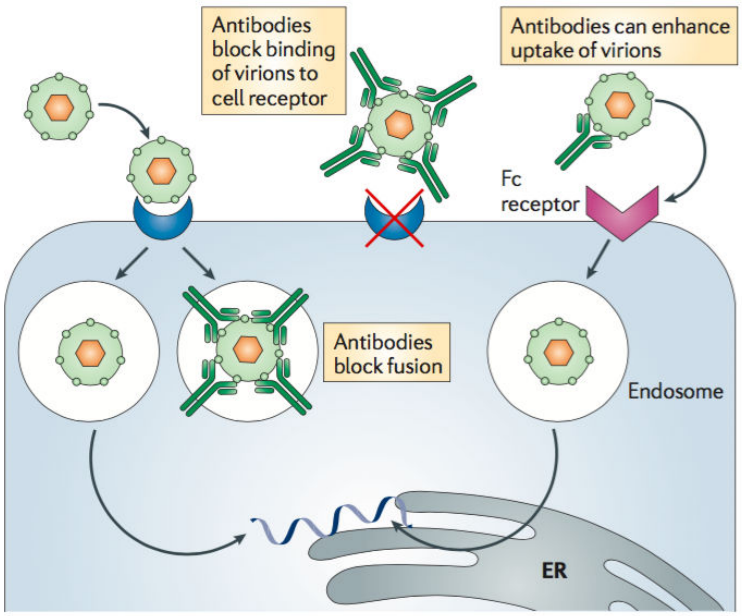


Figure 1.9: Antibody-dependent enhancement in secondary patients. Source: [Rothman, 2011].

1.2 Omics data types

1.2.1 Genomic data

General concepts for family and friends

The human genome consists of long macromolecules (chromosomes), sequences of nucleotides. Nucleotides can de distinguished by their bases. There are four types of nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Most of the time, in the nucleus of a cell, DNA is double-stranded. Strong covalent bonds bind bases together along a single strand, and weaker hydrogen bonds pair A with T and C with G between the two strands. Each single strand has two different ends called 5' and 3', oriented in opposite directions.

In gene-coding regions, the parts of the sequence, known as exons, are transcribed into RNA molecules that are, in turn, translated into proteins. The role of introns (the chromosomal regions that are within gene-coding regions but are not transcribed into RNA) and non-

gene coding (or intergenic) regions is only partially known and consists of a wide variety of regulatory elements for diverse functions. Changes in DNA sequence, either in exons, introns, or intergenic regions can lead to changes in the protein amino acids, or in their concentration, and thus affect human health, and reactions to pathogens.

Among the different types of DNA variation, we will here study single-nucleotide polymorphisms (SNPs). A SNP is a variation in a single nucleotide at a given position in the DNA (Figure 1.10) that occurs "quite often" in the population [Scitable by Nature Education, 2014]. There is no consensus on the precise frequency threshold, but it is usually on the order of one percent.

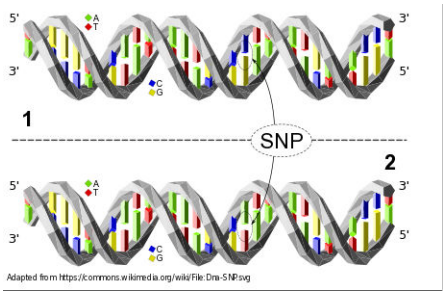


Figure 1.10: Single Nucleotide Polymorphism (SNP) within a DNA molecule. The two DNA molecules differ at a single base-pair location (a C/A polymorphism).

As the set of all, or most, SNPs, can be efficiently profiled using microchips, it is common to analyse genetic predisposition to different forms of disease, such as severe dengue, in that part of human genetic variation.

Genome-wide association study (GWAS)

The aim of genome-wide association studies (GWAS) is to find genetic predispositions to a given phenotype. Given two groups of samples from individuals with distinct phenotypes (e.g., forms of disease), a GWAS aims to identify SNPs for which the observed alleles are statistically associated with the different phenotypes.

For each sequenced SNP, one counts the number of occurrences of each SNP in cases and

in controls. Then, for each SNP, a statistical test assesses whether the allele counts in the two groups are significantly different. If they are, the SNP is said to be associated with the disease.

1.2.2 Gene expression data

Gene expression data represents the total amounts of distinct RNA transcripts in a cell. The entirety of RNA in a cell is called the transcriptome. While different types of RNA can be measured using transcriptomic technologies, we focus here on the measurement of messenger RNA (mRNA) concentrations which are often used as a proxy for protein concentrations in modeling, and therefore give a closer representation to activated/inactivated processes in cells. Gene expression is regulated by some genomic loci, known as expression quantitative trait loci (eQTLs). They can be situated within several hundreds of base pairs upstream or downstream of the gene region coding for the mRNA (cis-eQTLs), or elsewhere (trans-eQTLS). Gene expression is also regulated by the environmental factors such as disease state, immune history, diet, lifestyle such as smoking, pollution, etc., and changes over time, and between tissue types. Studying gene expression therefore enable to "integrate" environmental and genetic effects, and to therefore better explain, and understand resulting higher-level phenotypes.

1.3 Network analysis for biological data

By "network" we here mean a graph where nodes are genes, or proteins, for which these genes code. Edges are interactions between genes that were curated from sources independent of our disease-specific data: protein-protein interaction experiments such as yeast-to-hybrid, literature-curated interactions, experimental data from ChiP-chip experiments, co-expression data, etc. Edges are typically weighted, based on the nature of the data and the quantity of independent sources. Examples of such networks include STRING [Szklarczyk et al., 2014], I2D [Brown and Jurisica, 2007], HPRD [Peri et al., 2003], HumanNet [Lee

et al., 2011], and vary according to types of data included, more or less automated curation and size. Here, we mainly use STRING, since it is one of the broadest, most frequently updated, and well-documented, databases.

These networks have been shown to contain information about protein functions. This is due to the modular architecture underlying the molecular machinery of living systems [Barabási and Oltvai, 2004], composed of proteins that form relatively static complexes, such as the ribosome, as well as dynamically changing complexes such as immune complexes during infection.

The "guilt by association" principle states that proteins sharing common properties are likely to have similar functions and is commonly used in computational methods for protein function prediction. Previously, such methods were mainly based on information derived from proteins biochemical properties, their sequence [Friedberg, 2006] as well as their structure [Domingues and Lengauer, 2007]. By defining similarity measures on such properties, annotated proteins similar to a protein of interest can be found, and machine learning methods can be used to decide whether their functional annotations can be transferred (as e.g., in [Weinhold et al., 2008]). The "guilt by association" principle has, however, also been extended to predict protein function through proximity in protein interaction networks. Two main principles can be distinguished here: direct methods that use functional annotations enriched in the network neighborhood around a protein of interest, and module-assisted methods, which first identify modules of related proteins, typically by applying clustering approaches, and then annotating each module based on the known functions of its members [Sharan et al., 2007]. Large-scale network data has been proven useful not only to the functional annotation of proteins. A large number of computational approaches are guided by network data of different kinds, and in various ways.

1.4 Machine learning methods for biological data

1.4.1 What is a machine learning algorithm?

Definition

One of the first definitions was given by Arthur Samuel in 1959. According to him, machine learning gives "computers the ability to learn without being explicitly programmed." A more precise definition that was given in 1998 says that "machine learning explores the study and construction of algorithms that can learn from and make predictions on data" [Kohavi and Provost, 1998]. In other words, machine learning algorithms try to find patterns in existing data that would generalise to new incoming data.

Types of algorithms, based on input data

We can subdivide machine learning algorithms into three categories based on the input: supervised, unsupervised, and semi-supervised learning. Supervised learning requires "labelled" data, i.e., data for which we have input variable and already know the outcome. Its aim is to learn the relationships between the input variable and the outcome to be able to predict the outcome for new, "unlabelled" data. A general schema of supervised learning is presented in Figure 1.11.

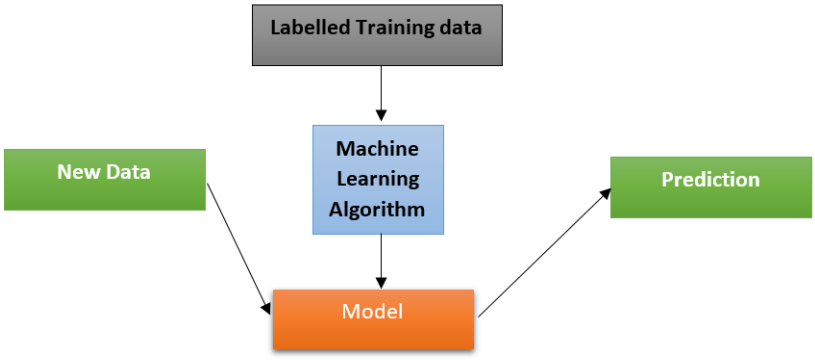


Figure 1.11: Supervised learning workflow. Source: https://www.codeproject.com

Supervised learning is often used to predict the phenotype of a patient, based on comprehensive molecular measurements, such as their genome, transcriptome, metabolome, etc. Another example is the prediction of patient phenotype in reaction to a viral infection. Unsupervised machine learning does not require the knowledge of any labels in advance. Clustering is a commonly used form of unsupervised learning. Finally, semi-supervised algorithms require a dataset with some known outcomes, and some (often many) unknown ones.

Generally, outcomes can be of different types: they can be continuous values (for instance, expression levels, protein levels, viral load...), or categories (type of disease, severity of disease...). In this thesis, I focus on supervised machine learning methods that can be used for classification.

Performance evaluation methods and terminology

The machine learning field uses some conventional names for different datasets used (*cf.* Figure 1.12). The initial data used to identify a model is called the "training data". Often,

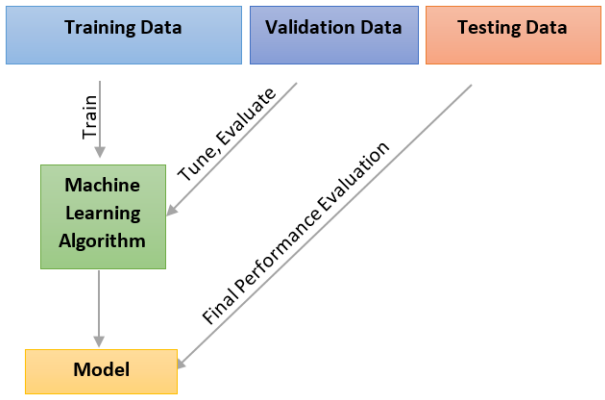


Figure 1.12: Supervised machine. Source: https://www.codeproject.com

additional data is required to infer additional parameters. This is the "validation data". The "test data" is used to evaluate the quality of the prediction on new data without using the previously learned patterns and parameters; only predictions are made.

When we have limited data, it is sometimes worth "mimicking" new samples using a technique called leave-k-out cross-validation (Figure 1.13). Leave-k-out cross-validation consists of iteratively leaving out k elements from the training data to keep them for future evaluation. The algorithm learns the model on the remaining elements, and then the performance is evaluated on the previously left out k elements. This procedure is then repeated with a new set of k elements. The number of iterations is typically chosen by the user. An advantage of leave-k-out cross-validation is that its result is based on the entire data, and not just one learning set. By the same token, test data is not overall independent from learning data; therefore, the variance of the cross-validation estimator can be large [Efron and Tibshirani, 1997]. For this reason, the comparison of models based on the results of cross-validation has limited value. The design of our evaluation procedures in Chapter 5 take this into account.

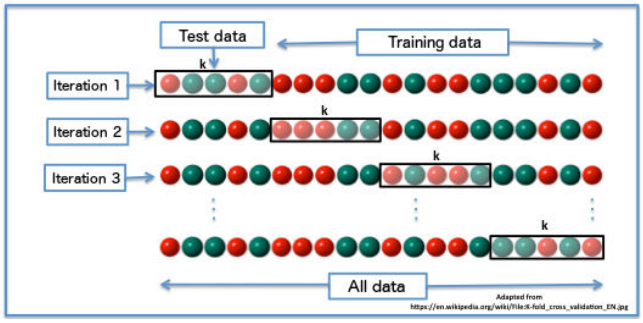


Figure 1.13: Leave-k-out cross-validation.

Choosing the right method: The bias-variance trade-off

The bias-variance tradeoff is a central problem in supervised learning. Ideally, one wants to choose a model to fit the data closely enough to capture its characteristic structure, but not too closely to avoid capturing the structure of the noise that is specific to the training sample ("overfitting").

Bias is the error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs ("underfitting"). This is the case of Model 1 in Figure 1.14.

Variance is the error from sensitivity to small fluctuations in the training set. High variance can cause overfitting: modeling the random noise in the training data, rather than the intended outputs. This is the case of Model 3 in Figure 1.14.

Ideally, one chooses a model that is "complex enough" to capture the characteristics of the data, i.e., the model is general enough to avoid erroneous assumptions (bias). On the other hand, the model should not be "too complex", i.e. the model assumptions should be specific enough to avoid sensitivity to small fluctuations in the data (variance). This is the case of Model 2 in Figure 1.14.

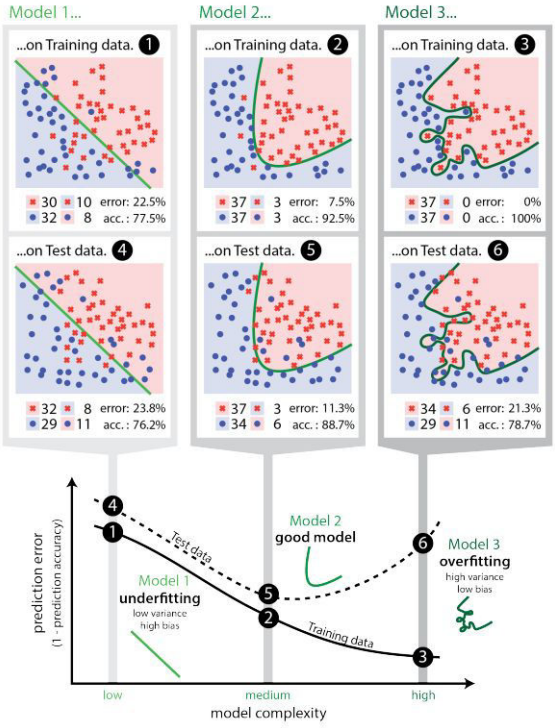


Figure 1.14: Bias-Variance trade-off. Source: https://cambridgecoding.wordpress.com

1.4.2 Mathematical framework and terminology

In this section, let p be the number of input features per sample (for instance, the number of transcripts per individual). Let $x_i \in \mathbb{R}^p$ be the i-th input. Let N be the total number of samples (for instance, the total number of patients in our case).

Let $X = (x_1^T, ..., x_i^T, ...x_N^T)$, $X \in \mathbb{R}^{N \times p}$ be the matrix of all inputs.

We denote by $|S|$ the size (or cardinality) of any set S.

Let $Y = \{0, 1, ..., C\}$, with $C = |Y| - 1$, be the finite set of possible classes that can be associated with any $x \in X$. Y can correspond to patient phenotypes. Let $y_i \in Y$ be the class of patient i. Let $\hat{y}_i \in Y$ be the predicted class of x_i.

1.4.3 Machine learning algorithms for supervised classification

To present a broad overview of the field, we here describe algorithms representing main approaches for the analysis of omics data. We focus on algorithms that are adapted to datasets where the number of features is larger than the number of individuals, as it is the case for our datasets. All presented methods are used for comparison with a newly designed method in Chapter 5. Explanations in this introductory part are adapted from [Hastie et al., 2009].

Instance-based learning: k-Nearest Neighbor (k-NN)

Instance-based learning is a family of learning algorithms that, instead of performing explicit generalization, compare new problem instances with instances seen in training. The most commonly used algorithm is this family is k-NN (short for "k-Nearest-Neighbour"). This algorithm is among the simplest of all machine learning algorithms.

k-NN finds the k closest training examples to an input sample using some predefined metric (such as Euclidean distance). The class of any input is then predicted to be the most common class among its k nearest neighbors (k is a positive integer, typically small, parameter chosen by the user). Figure 1.15 gives an illustration of a classification by such an algorithm.

The method of k-nearest neighbors makes very mild structural assumptions: its predictions are often accurate, but can be unstable, depending on the value of the parameter k.

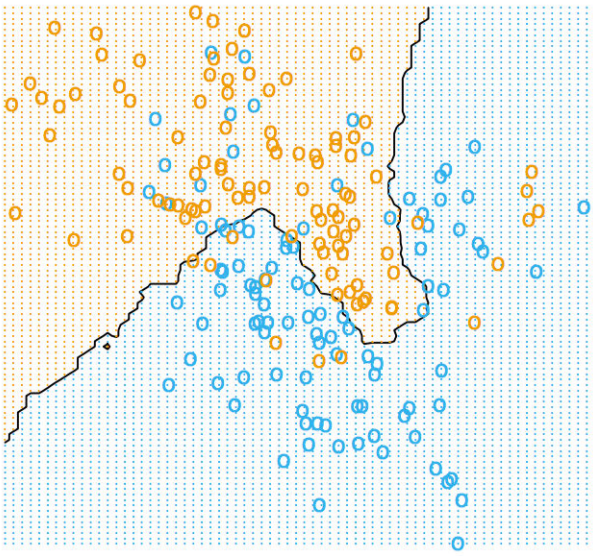


Figure 1.15: A classification example in two dimensions using k-NN classifier ($k = 15$). The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then fit by k-NN algorithm. The orange shaded region denotes that part of input space classified as ORANGE, while the blue region is classified as BLUE. Source: [Hastie et al., 2009]

Linear regression

Linear classification models are a classical, and still popular, choice. They make a very strong assumption regarding the relationship between input variables and classes. Linear models are simple and have relatively few parameters, thus being less prone to overfitting when $N \ll p$.

Given new matrix of inputs X, the output class vector $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_N)$ is predicted by the equation:

$$\hat{Y} = X\hat{\beta} + \hat{\beta}_0\vec{1}_N \tag{1.1}$$

where $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ is a vector of estimated coefficients and $\hat{\beta}_0$ corresponds to the constant coefficient, or the intercept at the origin, and $\vec{1}_N = (1, 1, \ldots, 1) \in \mathbb{R}^N$ is a vector of all ones of size N.

To avoid this additional constant in the above equation, we can integrate $\hat{\beta}_0$ into the product by replacing $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ by $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ and the input matrix $X =$

$(x_1^T, x_2, ..., x_N^T) \in \mathbb{R}^{N \times p}$, were $\forall i \in \{1 \dots N\}, x_i = (x_{1i}, \dots, x_{pi})$, by $X' = (x_1'^T, x_2'^T, ..., x_N'^T) \in \mathbb{R}^{N \times (p+1)}$, were $\forall i \in \{1 \dots N\} : x_i' = (1, x_{1i}, \dots, x_{pi})$. For the sake of simplicity, we will not change the notations X to X' and β to β' in the following, but the constant will be included in the input variables.

With this change in notations, the Equation 1.1 can then be rewritten as:

$$\hat{Y} = X\hat{\beta} \tag{1.2}$$

In the case of supervised classification, we have a set of patients for whom we know the class (i.e., the training set). From this set we would like to estimate all the β_i coefficients by minimising an error between the real phenotypes of our training set patients and their predicted phenotype, using the linear model. To quantify the error, different metrics can be chosen. The most commonly used method, known as the method of *least squares*, consists in minimising the residual sum of squares (RSS):

$$RSS(\beta) = \sum_{i=0}^{N} (y_i - x_i^T \beta)^2.$$

By taking the derivative and searching for the point at which the derivative is equal to 0, we find the formula of the extremum (that is a minimum, given the fact that $RSS(\beta)$ is a sum of squares, thus has a quadratic form and stays positive):

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

where $(X^T X)^{-1}$ is the pseudo-inverse of $X^T X$.

Therefore, this method provides an analytic expression of a the optimal coefficients. (For matrices where $N \ll p$, $X^T X$ is singular, multiple optima exist.)

The above estimation is unbiased, i.e., for an infinite number of inputs, we obtain that the expected value of $\hat{\beta}$ is β: $\mathbb{E}(\hat{\beta}) = \beta$. Once the coefficients are estimated, we determine the separation between classes. The separation between two classes corresponds to points, where the assignment to any of the two classes generates the same error. For linear regression, this corresponds to points where $\hat{Y}$ is constant and equal to some threshold th, between the

two classes (typically for classes 0 and 1, $th = 0.5$):

$$th = X\hat{\beta}.$$

This is an equation of a hyperplane, thus linear regression always separates classes by a hyperplane. Figure 1.16 illustrates the result of a classification by linear regression.

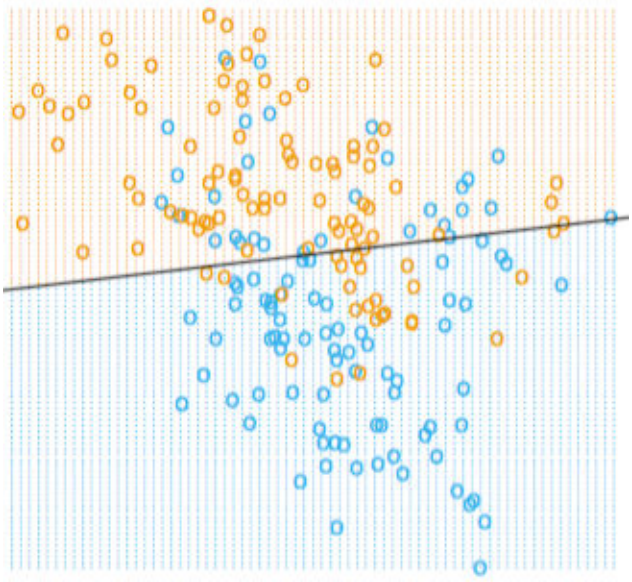


Figure 1.16: A classification example in two dimensions using linear regression. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then fit by linear regression. The orange shaded region denotes that part of input space classified as ORANGE, while the blue region is classified as BLUE. Source: [Hastie et al., 2009]

Linear regression with least squares estimates of coefficients is the simplest model-based regression approach. It is well suited for small datasets, because it has relatively few degrees of freedom. Moreover, the linear model is quite intuitive to interpret.

The drawback of this method is that the result can be inaccurate whenever the underlying true relationship between input and output is not linear.

Feature selection using Lasso

Even when the relationship between input and output is linear, straightforward linear regression may be problematic, for two reasons that are important for learning from omics data:

The first reason is prediction accuracy. In particular for the case of many features, least squares estimates often have low bias but large variance.

The second reason is interpretation. Instead of large models with a many features, one often would prefer smaller, more easily interpretable, subset of variables that exhibits the strongest effect on the output.

Lower variance and a lower number of features are typically achieves by incorporating a feature selection penalty into the optimisation objective. There are three common approaches: Ridge regression, Lasso, and Elastic Net. Here, we will present the Lasso method, since it is the one that generates the sparsest solutions.

In the Lasso approach, one optimizes the coefficients as in least squares, but imposes a bound on the so-called Lasso penalty:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t.$$

This minimisation problem subject to a constraint may be rewritten using the Lagrangian function as:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where λ is a parameter that depends on the choice of t. In practice, the parameter λ is optimised to minimise misclassification error when performing leave-out cross-validation on the learning dataset.

Because of the nature of the constraint, making t sufficiently small (or, equivalently, λ sufficiently large) will cause some of the coefficients to be exactly zero. Therefore, less features will be used to predict the outcome. As a consequence, the prediction is slightly biased, but the variance of the predicted values will decrease, and the set of features becomes easier to interpret.

Logistic regression

Logistic regression is an adaptation of linear regression that is better suited to classify data with a limited number of output classes (it is especially suited for binary classification, i.e., where we only have two classes 0 and 1).

Logistic regression applies a logistic function to a linear combination of the input variables before learning parameters for classification. The logistic function $\sigma(t)$ is:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

With $t = X\beta$, we get the logistic regression function,

$$\sigma(X) = \frac{1}{1 + e^{-X\beta}}$$

Figure 1.17 illustrates a logistic function in two dimensions.

On the example plot, a point x_i with input value lower than -4 and class $y_i = 0$ will be well fit by this regression (i.e., $\hat{y}_i$ will be close to y_i), even in the presence of a point x_j with an input value $\ll 0$ and output $y_j = 0$. This would not be the case with a linear fit. Similarly, an input value x_l greater than 4 and class $y_l = 1$ will be well fit by such a regression. With a better fit, the resulting classification can be expected to perform better. The logistic function arises from the objective to model the posterior probabilities of our classes via linear functions in X, while ensuring that they sum to one and remain in [0,1]. Details can be found in [Hastie et al., 2009].

Importantly, when using logistic regression for classification, the separation between classes remains linear. Similarly to linear regression, the separation between two classes satisfies

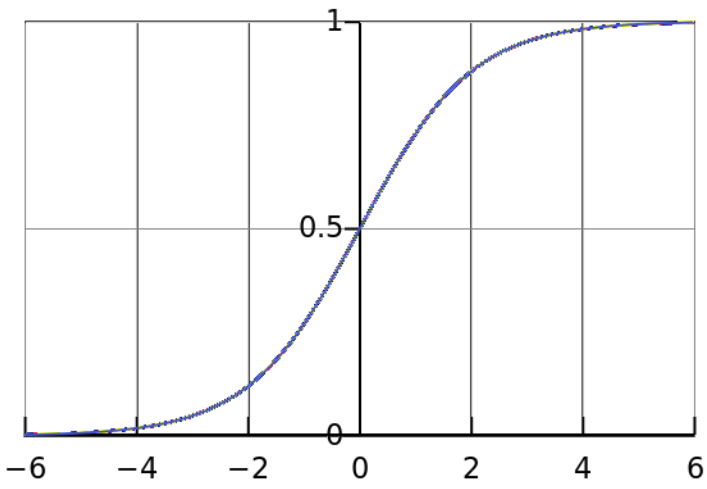


Figure 1.17: The logistic function $\sigma(t) = \frac{1}{1+e^{-t}}$

for some threshold th : $th = \frac{1}{1+e^{-X\beta}}$. Since the logistic function σ is monotonic, this is equivalent to:

$$\sigma^{-1}(th) = \ln\left(\frac{th}{th-1}\right) = X\beta,$$

where σ^{-1} is the inverse function of the logistic function. This corresponds again to an equation of a hyperplane.

The logistic function thus adapts linear regression for binary classification. Just as for linear regression, we can combine logistic regression with a feature selection penalty to improve accuracy and interpretability and adding some bias. Nevertheless this approach still fits a very specific function to the data. If the data does not follow this function, bias and inaccurate predictions result.

Linear Support Vector Machines (linear SVMs)

Support vector machines are a family of methods that can be used for supervised classification, and not based on regression. I will here present linear SVMs for binary classification.

The basic idea of SVMs is to find a linear boundary (a hyperplane) that not only minimises

the number of misclassified points, but also aims to be as far as possible from any point in any class. For a dataset that can be perfectly separated by a hyperplane, the algorithm will maximise the size of a margin between the separation and the closest point on each side of the boundary. The right panel of Figure 1.18 illustrates this problem.

For the cases in which it is not possible to perfectly separate the two classes, a penalty is included for the misclassified individuals. This penalty is proportional to the distance to the margin. This case is illustrated on the left panel of Figure 1.18.

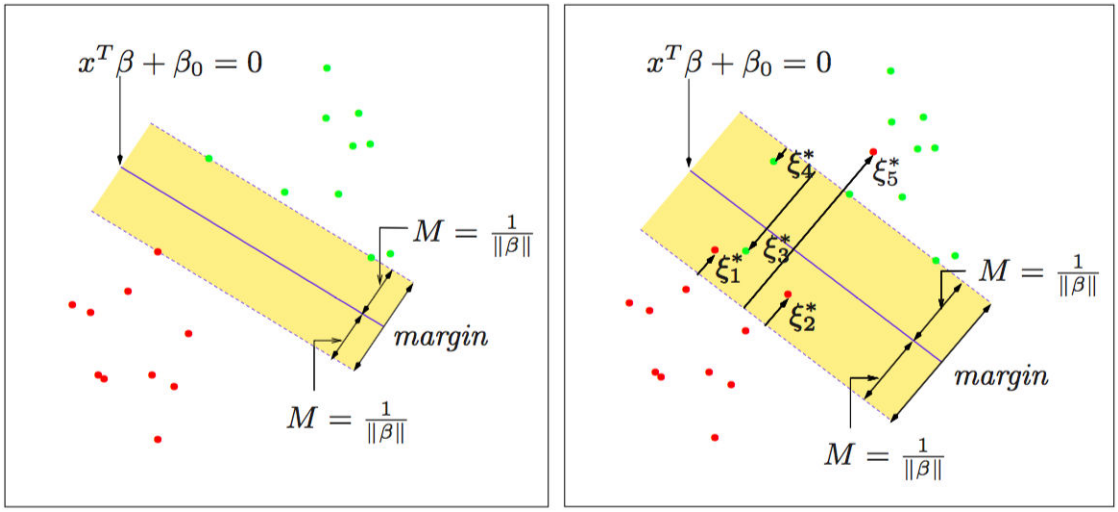


Figure 1.18: Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = \frac{2}{||\beta||}$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = M\xi_j$; points on the correct side have $\xi_j^* = 0$. The margin is maximized subject to a total budget $\sum \xi_j \leq constant$. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin. Source: [Hastie et al., 2009]

For mathematical simplicity, we will assume in this section that our classes y_i are -1 or 1: $\forall i \in \{1, ..., N\}, (x_i, y_i)$ are the pairs of input variable and class of the individual i, where $y_i \in \{-1, 1\}$ and $x_i \in \mathbb{R}^p$. In the case where we can find a perfect boundary, we want to solve:

$$\max_{\beta,\beta_0,||\beta||=1} M \tag{1.3}$$

subject to $\forall i \in \{1, ..., N\} : y_i(x_i^T\beta + \beta_0) \geq M$. This problem can be rewritten without

explicitly mentioning the margin M. By relaxing the constraint $||\beta|| = 1$ and setting $M = 1/||\beta||$, we can show that an equivalent formulation is:

$$\min_{\beta,\beta_0} ||\beta||, \text{subject to } \forall i \in \{1, ..., N\} : y_i(x_i^T\beta + \beta_0) \geq 1. \tag{1.4}$$

This is the usual way of writing the support vector criterion for the case where all points of the learning set can be correctly classified by the learned model.

For the case where we cannot find a boundary that perfectly classifies every element, the SVM problem in Expression 1.3 is adapted using the following constraint:

$$\forall i \in \{1, ..., N\} : y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i),$$

subject to $\xi_i > 0, \sum \xi_i \leq constant.$

The equivalent formulation to Expression 1.4 becomes:

$$\min ||\beta||, \text{ subject to } \forall i \in \{1, ..., N\} : \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\ \xi_i > 0, \sum \xi_i \leq constant. \end{cases}.$$

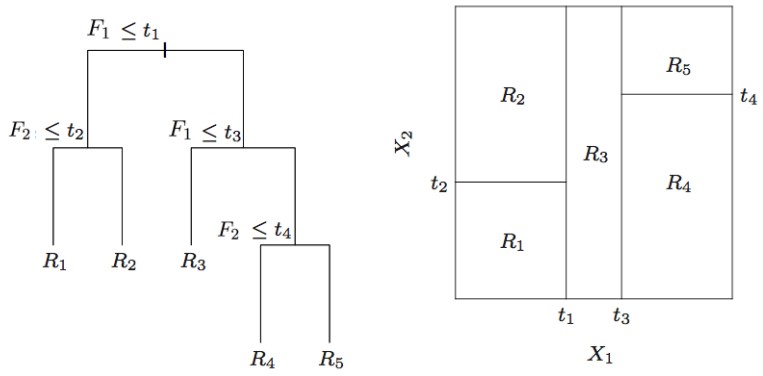
This is the usual way the support vector classifier is defined for the non-separable case. This makes SVM a good approach to find linear boundaries to classify data. Specific types of nonlinear boundaries may also be constructed by applying a transformation (known as kernel), satisfying specific properties, to the original features, and determining a linear boundary within this transformed space. Kernels were not used within the scope of this work.

Decision trees

In this section, let $X = (F_1, ...F_p)^T, F_i \in \mathbb{R}^N$, F_i being a vector of values of feature i for N samples. Tree-based methods provide a conceptually simple way to learn non-linear boundaries. Tree-based learning methods work recursively: Given a learning set, they search for a feature F_i that separates optimally (according to some criterion) samples from different classes using a threshold t_i. For each corresponding subset, they again search for

the feature F_j that separates best different classes within that set using a threshold t_j, and so on, until some stop criterion is satisfied (involving, for instance tree depth, or best cross-validation performance). The class assigned to a leaf is typically the class to which a majority of the set of samples from this leaf belongs to.

Decision trees are typically represented as rooted binary trees. Each internal node represents a single input variable and a split point on that variable. The leaf nodes of the tree contain the output class. For a new sample, class prediction is performed by walking down a path of the tree starting from its root, iteratively following branches according to the learned split points, and outputting the class value at the leaf node (Figure 1.19).



(a) Decision tree **(b)** Corresponding space partition

Figure 1.19: Decision tree approach and resulting partition of the feature space. Here, the set of classes is $Y = \{R_1 \dots R_5\}$, and the feature vectors are F_1 and F_2. Adapted from [Hastie et al., 2009].

Trees are fast to learn and very fast for making predictions. A weak point of decision trees is that they have a relatively high variance, and generally overfit more than other methods.

Random forests

Random forests are an ensemble approach to decision trees that aims to decrease variance/overfitting. To do so, during the learning process, it applies a technique called "bag-

ging"(short for "bootstrap aggregating"). Bagging repeatedly (B times) selects, and fits a decision tree to, a random sample of inputs x_i:

For $b = 1, ..., B$:

- Sample, with replacement, B training samples from the inputs X and their classes Y; call these X_b, Y_b.
- Train a decision tree f_b on X_b, Y_b.

The output consists of B decision trees. New predictions for unseen samples x' can be made by taking the class most frequently attributed to x' by the B decision trees (breaking ties where necessary).

The number of samples/trees, B, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees B can be found using cross-validation.

Random forests differ in only one way from bagging: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features, from which the feature to define the split is selected. This process is sometimes called "feature bagging".

Typically, for a classification problem with p features, $\sqrt{p}$ (rounded down) features are used in each split.

Random forests provide an adaptable way to search for non-linear boundaries with a very adaptable model. They often have better predictive accuracy than decision trees, but the interpretation of the learnt feature is usually extremely difficult, due to its complex structure.

Chapter 2

Network analysis to aggregate dengue genotyping data

Contents

2.1 Introduction

Epidemiological studies have repeatedly shown that severe dengue is associated with ethnicity [Coffey et al., 2009, Bravo et al., 1987, Guzman and Kouri, 2002, Halstead et al., 2001]. The research community has hypothesised that this is due, in part, to the genetical background. To date, several genomic associations with severe dengue have been identified using GWAS analysis [Khor et al., 2011, Whitehorn et al., 2013]. Nevertheless, the research community has trouble linking these associations with disease etiology. One hypothesis to explain this difficulty is that dengue is a complex disease, i.e., influenced by a combination of multiple genes. If we search for associations between one of the genomic positions and the phenotype in the whole genome and independently from one another, we will be testing the same hypothesis many times, and will need to correct results for multiple testing. When the genomic positions are hundreds of thousands of SNPs, as for most published dengue analyses, the correction will be very strong. Therefore, only the very strong associations would remain statistically significant, while many polymorphisms with small marginal effects will be undistinguishable from random noise [Eichler et al., 2010, Maher, 2008]. If we are able to correctly group the marginal effects in one signal, these effects may add up and become statistically distinguishable from random noise. For instance, we can aggregate SNP p-values by some known biological units such as genes. We may even then further group gene p-values by known common functions such as pathways. When there is a risk that the useful pathways are not entirely present in the databases, we may simply use the broader information about gene-gene interactions, and aggregate gene p-values by sets of interacting genes from gene interaction networks. This chapter will discuss my work on aggregating dengue GWAS data using available knowledge to identify genes or groups of genes associated to severe dengue. I will first describe how I aggregate SNP-level information into gene-level information, then apply existing pathway analysis, and finally apply gene interaction network analysis algorithms (*cf.* Figure 2.1).

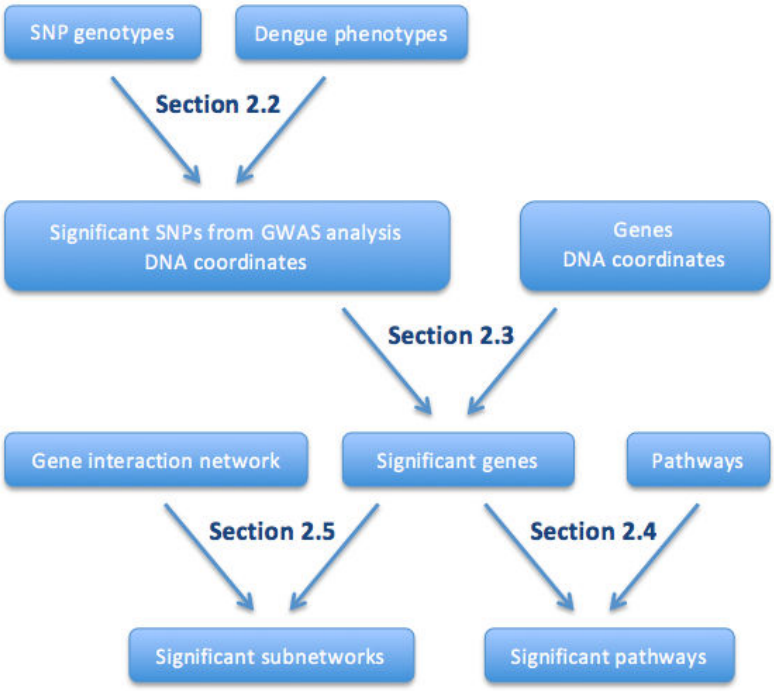


Figure 2.1: Summary of the analyses performed in this chapter.

2.2 Dataset

In this section, I analyse a case-control cohort from Vietnam whose GWAS was previously published [Khor et al., 2011]. It contains 2008 pediatric cases treated for dengue shock syndrome (DSS) and 2018 controls.

Cases were eligible if they were under 15 years of age and had clinical signs, symptoms and hematological findings that led to a clinical diagnosis of incipient or established DSS, as defined by the WHO 2009 report [WHO (World Health Organisation), 2009]. Are considered to be in shock those patients that show warning signs, and whose pulse pressure is lower than 20 mmHg, or showing signs of poor capillary perfusion (cold extremities, delayed capillary refill, or rapid pulse rate). Blood samples for research and diagnostic tests were collected at the time of enrolment and again before patient discharge from hospital. Patients enrolled were recruited in the pediatric intensive care unit of the Hospital for Tropical Diseases (Ho

Chi Minh City, Vietnam) between 2001 and 2009. Parents or guardians of each participant gave written informed consent to participate. The Scientific and Ethical Committees of each study site approved the study protocols, as did the Oxford University Tropical Research Ethical Committee.

Variable	N(%) or Median (Quartiles)
Age	9 (6-11)
Male	1036 (51,5%)
Day of illness [a]	5 (5-6)
Established DSS symptoms	
* Hypotensive, rapid or weak or absent pulse, +/- poor peripheral perfusion	647 (32%)
* Narrowed pulse pressure (≤20mm/Hg), rapid or weak pulse and poor peripheral perfusion	1023 (51%)
Incipient DSS symptoms	
* Narrowed pulse pressure (≤20mm/Hg), rapid or weak pulse, perfused periphery	224 (11%)
* Pulse pressure=25mm/Hg, rapid or weak pulse, +/- poor peripheral perfusion	115 (6%)
* Pulse pressure>25mm/Hg, +/- rapid or weak pulse, +/- poor peripheral perfusion	0
Maximum Hematocrit [a]	49 (46-52)
Platelet nadir [a]	35 (22-53)
Spontaneous bleeding	1674 (83%)
Fatal outcome [b]	5 (0.2%)
Dengue virus serotypes	
* DENV-1	971 (48%)
* DENV-2	233 (12%)
* DENV-3	34 (2%)
* DENV-4	60 (3%)
* Mixed infection	2 (0.1%)
* Negative RT-PCR	345 (17%)
* Unknown	364 (18%)

Figure 2.2: Characteristics of the patients of the GWAS cohort. a: missing information for five patients. b: missing information for one patient. Reproduced from [Grange, 2014].

Controls consisted of sequenced cord blood samples, and were from newborns. They were collected at Hung Vuong Hospital (Ho Chi Minh City, Vietnam) between 2004 and 2006. All participants gave written informed consent to participate. The Scientific and Ethical Committees of each study site approved the study protocols, as did the Oxford University Tropical Research Ethical Committee. DNA was extracted from cord blood using Nucleon BACC Genomic DNA Extraction Kits (GE Healthcare, USA).

Genotyping was performed with Illumina Human 660W Quad BeadChips following the manufacturers instructions. Cases and controls were randomized on plates and genotyped. Out of the initial 500,000 SNPs, 428,910 remained after quality control. The quality control criteria excluded: SNPs that had genotypes with more than 5% missing, showed gross departure from Hardy-Weinberg equilibrium (a departure with a p-value $\leq 10^{-7}$), or had

a minor allele frequency below 5%. For sample quality control, samples with an overall genotyping call rate of lower than 95% were excluded from analysis. SNPs that had a p-value that was lower than the Bonferroni-corrected threshold p-value: $\alpha_{Bonf} = \alpha/n_{SNPs} = 0.05/428910 = 1.2 * 10^{-7}$ were considered as significant. To check the coherence of the data, I reran the GWAS analysis published in [Khor et al., 2011] using PLINK (v1.7) [Purcell et al., 2007]. Figure 2.3 is the resulting Manhattan plot. Two distinct regions reach the Bonferroni-corrected significance threshold, as described in the initial paper.

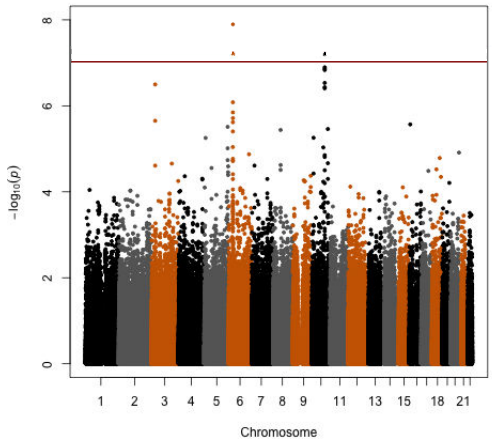


Figure 2.3: Manhattan plot of the replicated GWAS results. The horizontal axis represents the SNPs ordered by chromosomes and chromosomal positions, the vertical axis represents $-\log$ of SNP p-values, under the null hypothesis of no association. The red line represents α_{Bonf}, the significance threshold of $\alpha = 0.05$ corrected for multiple testing using the Bonferroni correction.

2.3 Aggregating genomic information to the gene level

Since dengue is a complex disease, we wonder whether more can be learnt from this genomic data, than the associations of only two genes to severe dengue. In the previous analysis the Bonferroni-corrected threshold for statistical significance is very stringent, because we test the hypothesis of association to severe dengue 428,910 times (one for each SNP). If we were able to correctly group the less strong associations in one signal, these effects together may

become statistically significant. In this chapter, we will aggregate SNP p-values by genes, and analyse the resulting p-values.

2.3.1 Methods

Mapping by genomic position

As long as we do not have a map of all interactions and regulatory relationships between DNA nucleotides, aggregating SNPs into some functional units needs to been done heuristically. The most direct way to map SNPs to genes is to identify SNP and gene location on the DNA chromosomes and annotate which SNP is included in which gene. I first downloaded genomic positions of genes from RefLink table of the UCSC database (http://genome.ucsc.edu/) [Karolchik et al., 2004]. From my mapping of SNP positions to genomic positions, it appears that 53 % of SNPs are located in intergenic regions (i.e., outside of known genes). If we analyse SNPs included in genes only, intergenic SNPs would thus be deleted from the analysis! Can we improve the mapping of intergenic SNPs to genes to avoid losing more than half of the SNP information in downstream analysis?

A commonly used heuristic to go beyond the "straightforward" form of mapping consists of mapping to genes SNPs that are in the flanking regions [Liu et al., 2017]. Indeed, it is assumed that these regions are enriched in binding sites of regulatory elements such as promoters, transcription factors etc. DNA modifications at binding sites may impact the binding affinity of the regulatory element of interest, therefore affecting the regulation of the expression of the corresponding gene. Thus, it is usually deemed useful to map SNPs within a few kilobases (kb) to the left and to the right of the gene to the gene of interest. In this analysis, the size of the flanking region is 10kb.

Tissue type	Patients	cis-eQTLs	trans-eQTLs	Multiple testing correction	Reference
Whole blood	5,311 + 2,775	585,669	1,152	1% FDR	[Westra et al., 2013]
Skin	847	2,796	609	1% (cis), 10% (trans) FDR	[Grundberg et al., 2012]
Adipose tissue	855	3,529	639	1% (cis), 10% (trans) FDR	[Grundberg et al., 2012]
Liver	427	1,350	491	10% FDR	GTEx [Schadt et al., 2008]
LCL*	1,355	6,579**	11,977**	5% FDR	[Liang et al., 2013]

Table 2.1: Description of eQTL data sources useful for dengue analysis.

*: LCL stands for lymphoblastoid cell lines. These are immortalised cell lines of B cells.

**: number of genes that have at least one SNP that regulates their expression.

Exploration of available functional information

To improve the coverage of the mapping, I investigated the possibility to use functional information about gene regulation by SNPs, such as expression quantitative trait loci (eQTLs). Briefly, eQTLs are genetic regions that are statistically associated with modified levels of the expression of a specific gene (*cf.* Part 1.2.2). SNPs are known to be enriched in regulatory elements, such as eQTLs, relative to the rest of the genome [Cookson et al., 2009, Nicolae et al., 2010]. Statistically speaking, eQTL analysis aims at finding regulatory relationships between SNPs and gene expression modifications by searching for correlations between the expression level of a gene and SNP alleles. Therefore, such an analysis requires genotyping and gene expression information for the same patient. We did not have sufficient genomic and transcriptomic data from patients that would have enabled us to establish eQTLs for the South-Asian population. I thus searched for eQTLs that may be relevant for the reaction to dengue virus.

I surveyed datasets in the databases GTEx [Lonsdale et al., 2013], SCAN [Gamazon et al., 2010], eQTL uChicago [Veyrieras et al., 2008], SeeQTL [Xia et al., 2012] and the datasets of [Westra et al., 2013], [Liang et al., 2013]. Since eQTLs are related to gene expression, they are tissue-specific. Therefore, I focused on datasets of eQTLs related to tissues that are suspected to play a role in dengue etiology. Table 2.1 gives details on the largest datasets found in the above databases for each relevant tissue type.

Since our priority is to map the intergenic SNPs to genes, we are specifically interested in

trans-eQTLs (i.e., eQTLs that are not situated in the gene whose expression they regulate, as opposed to cis-eQTLs that are situated within the gene that they regulate). From Table 2.1, we can see that there is a limited number of trans-eQTLs that can be found in each dataset. Indeed, to find trans-eQTLs, one needs to test for association between every SNP of interest with every gene of interest. This results in many tests; therefore the corresponding multiple test correction strongly reduces power, and requires large sample sizes to allow many significant hits. Moreover, trans-eQTLs are rarely reproduced in other datasets [Liang et al., 2013], and even more so when they are calculated on different subpopulations. None of the eQTL databases I surveyed perform an analysis on populations of Asian origin, and thus matching our data on dengue. Our analysis may be particularly sensitive to genetic background, as the proportion of severe dengue cases is known to vary strongly in different parts of the world. Additionally, from a statistical point of view, SNPs may be associated with the expression of several genes. For instance, in the whole-blood dataset in Table 2.1, authors report 103 independent SNPs at the origin of all of the 1,152 trans-eQTLs found in the analysis. It means that a SNP would on average be mapped to ten genes! Thus, mapping them to these genes will add dependencies between gene p-values, and will require further aggregation of gene p-values to take into account these dependencies. Using eQTL results from different databases creates other challenges: experimental techniques vary, samples are of different sizes, different statistical tests have been used to find eQTLs and to correct for multiple testing, some results are adjusted for confounders but others are not, some are adjusted for batch effect but others are not, multiple testing corrections vary, etc.

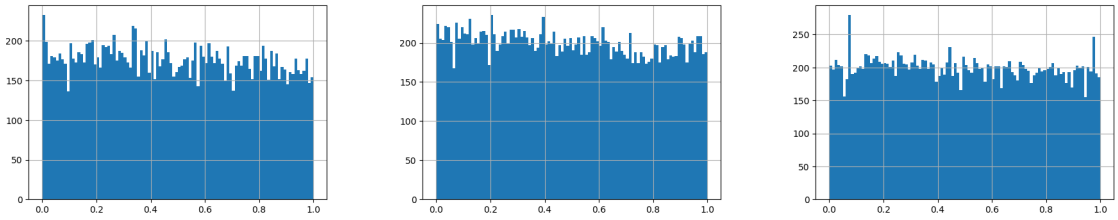
Since, in my analysis, the achievable advantage from integrating these results, and the achievable quality of the mapping outside of coding regions was not clear, I decided to continue with a simple physical mapping, as described in the following section.

Gene p-value computation

To combine SNP-level p-values obtained from GWAS into gene-level p-values, one needs to take varying gene lengths and to the potential statistical dependencies in between neighbouring SNP alleles, known as linkage disequilibrium (LD), into account. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly [Slatkin, 2008].

I used VEGAS [Liu et al., 2010], a tool that takes into account gene length and LD. VEGAS prunes SNPs that are in LD using a HapMap LD map, then aggregates p-values of gene SNPs of interest into a test statistic, and then an empirical p-value. VEGAS gives flexibility as to which SNPs to agglomerate into a gene-based p-value. Indeed, for some genes, an approach considering all SNPs might be the most powerful; for others, focusing on a certain percentage of most significant SNPs may be more powerful, for others only one most significant SNP carries all the information. The best methodology depends on the generally unknown proportion of SNPs in a gene that influence the underlying biological process of interest.

For my application, I tried different possible options: I aggregated all SNPs to genes, only the top SNP of each gene, or only the top 10% SNPs of each gene (this last option was rerun using the new version VEGAS2v02 [Mishra and Macgregor, 2017], where a statistical mistake was corrected [Hecker et al., 2017]). To determine LD structure, I used the Hapmap LD map of eastern Asian populations (i.e., HapMap Han Chinese in Beijing population and Japanese in Tokyo populations), since that was the closest to the population origin of the Vietnamese dataset. Figure 2.4 represents a histogram of the distribution of these p-values. Since the top SNP-based gene p-values where most enriched in low values, I carried on the analysis with these gene-level p-values.



(a) Top SNP p-values **(b)** Top 10% SNP p-values **(c)** All SNP p-values

Figure 2.4: Histograms of gene-level p-values when aggregating different subgroups of SNPs contained in each gene.

2.3.2 Results

SNPs were mapped to 17,629 genes, with five genes having Bonferroni-corrected p-values below the 0.05 threshold (Table 2.2). Interestingly, three out of the five genes, HLA-B, MICA, and HCP5, are all related to the major histocompatibility complex (MHC), also known as human leukocyte antigen (HLA) complex. The function of MHC molecules is to bind peptide fragments derived from pathogens, and to display them on the cell surface for recognition by the appropriate T cells. Consequences of mutations in these genes are almost always deleterious to the pathogen-infected cells who are killed; macrophages are activated to kill bacteria living in their intracellular vesicles, and B cells are activated to produce antibodies that eliminate or neutralise extracellular pathogens. Thus, there is strong selective pressure on this gene region. Indeed, the MHC is known to contain a high number of genetic variants of each gene within the population as a whole. The MHC genes are, in fact, the most polymorphic genes known [Janeway CA et al., 2001]. The evolution of these genes is thought to be driven by the differences in pathogens encountered by their hosts in the course of human evolution. This may explain differences in severe dengue susceptibility in different populations with different genetic background.

2.3.3 Discussion

The above results seem to suggest that differences in the MHC complex are related to genetic susceptibility to severe dengue. This result confirms some findings about dengue

Gene	Corrected p-value	NCBI Gene Description
NOC3L	<0.01	NOC3-like DNA replication regulator
PLCE1	<0.01	Associated with severe dengue in the GWAS of this data [Khor et al., 2011]. It encodes a phospholipase enzyme that catalyzes the hydrolysis of phosphatidylinositol-4,5-bisphosphate to generate two second messengers: inositol 1,4,5-triphosphate (IP3) and diacylglycerol (DAG). These second messengers subsequently regulate various processes affecting cell growth, differentiation, and gene expression. Mutations in this gene cause early-onset nephrotic syndrome.
HLA-B	0.01	Major histocompatibility complex (MHC), class I, B. Class I molecules play a central role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen. They are expressed in nearly all cells.
MICA	0.02	MHC class I polypeptide-related sequence A. The associated protein is highly polymorphic.
HCP5	0.02	MHC complex P5 (non-protein coding)

Table 2.2: Mapping SNPs to genes: Description of significant VEGAS genes

etiology [Stephens et al., 2002, Lan et al., 2008].

However, my analysis is limited by the required high number of ad-hoc choices made to map SNPs to genes. Indeed, all SNPs have been mapped to genes physically without including intergenic SNPs and mapping intronic SNPs to the gene they were in, ignoring any functional link to other genes. We had considered integrating eQTL information. From a biological point of view, this information is population- and tissue-specific. Since no eQTLs are available for the Asian population, mapping eQTLs of populations with European and African origin may generate many false positives. From a statistical point of view, it is very difficult to integrate datasets, since, typically, they use different batch corrections, statistical tests and multiple testing corrections. Moreover, since one SNP can control the expression of several genes, mapping a SNP to several genes introduces strong dependencies in between gene p-values that would need to be taken into account during network analysis,

which hinders the downstream statistics. Based on the little information available and the discussed disadvantages, we decided to not include this information for the SNP-to-gene mapping. In the future, if we wanted to improve the mapping, we could also consider using chromosome architecture information, since it is known that SNPs that are physically close to genes in a folded chromosome may influence the expression of that gene by making DNA more or less accessible for gene expression.

Once SNPs have been assigned to genes, there are also choices to be made as to how to map SNP p-values to genes. We have used a data-driven approach to choose the method generating the strongest statistical signal. In this case, mapping the top SNP to the gene appeared to be the best among the three tested options. Nevertheless, such a mapping relies on one SNP, and may thus be more prone to noise than the other mappings. I might have tried to test different percentages of SNPs to map to a gene in the analysis, but this might have led to overfitting. In reality, the proportion of SNPs that carry some association signal may vary not only from disease to disease, but also from gene to gene. To figure out the best mapping for each gene, a very large sample size would be needed.

2.4 Pathway analysis

One way to improve robustness and quantity of results is to include more functional regulation information and is to use pathway-based and network-based analyses that do not limit themselves to statistically significant genes, but aim to identify groups of genes that are functionally related and are enriched in low p-values. I first performed the more classical pathway analysis.

2.4.1 Methods

Several tools exist for pathway analysis. They differ by the input data type, enrichment statistic and by the pathway database they use to group genes into pathways (ConsensusPathDB [Kamburov et al., 2011], Ingenuity Pathway Analysis (Ingenuity Systems, GenGen

[Wang et al., 2010], Reactome [Fabregat et al., 2016]...). Among them, GSEA [Subramanian et al., 2005] is widely used. GSEA was originally created to assess gene set enrichment in transcriptome data.

It uses the Molecular Signatures Database (MSigDB) to define gene sets. MSigDB is a compilation of collections of annotated gene sets that includes main pathway databases, along with other more specific collections of gene sets derived from the literature. Each collection of gene sets can be used as a *background dataset* for enrichment analysis. The background dataset defines gene sets and quantifies the proportion of genes belonging to each gene set in the whole genome. When given a list of input genes sorted by any score, GSEA tests the null hypothesis of whether the top (or the bottom) of the gene list is enriched in genes from some of the defined gene sets, compared to the background dataset. The output is a q-value of such an enrichment for each gene set. A q-value is the lowest FDR threshold at which the result becomes significant. In other words, a genes set with a q-value q will be considered as significant if and only if we accept to have a proportion q of results being false positives. I have used GSEA to search for enriched pathways using diverse background datasets:

1. A "hallmark" gene set that contains gene sets derived by aggregating many MSigDB gene sets to represent well-defined general biological states or processes.

2. An immunology-specific gene set containing genes differentially expressed under different stimuli (reaction to different pathogens, or to molecules activating immunity).

3. KEGG dataset [Kanehisa and Goto, 2000]. KEGG is a database of manually curated and represented pathway maps summarising the current knowledge on the molecular interactions. It is broadly used and frequently updated.

4. Reactome dataset [Croft et al., 2011]. Reactome is another manually curated database that represents pathways. But the unit of the Reactome data model is the reaction. Interacting entities are diverse: nucleic acids, proteins, complexes, vaccines, anti-cancer therapeutics, and small molecules. Reactions are grouped into a network, and

then, pathways.

5. An aggregation of curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts available via GSEA.

I have then used a commercial pathway enrichment tool, Ingenuity Pathway Analysis (Ingenuity Systems, http://www.ingenuity.com), that has a hand-curated database of widely recognised and high quality pathways.

To avoid bias from the step of grouping SNPs to genes, I also used VEGAS2 Pathway [Mishra and MacGregor, 2017], which does not rely on a grouping of SNPs by genes, but directly groups them by pathways.

2.4.2 Results

At an FDR threshold of 20%, none of my analyses have led to the detection of an enrichment.

2.4.3 Discussion

Our negative results may indicate that the statistical association at the level of pathways is not strong enough to be significant. One reason for that may be that, in a disease whose etiology is still largely unknown, many relevant pathways still need to be discovered, or represented in pathway databases. Additionally, even in known pathways, only genes in a small part of a large pathway may be associated with the disease (a pathway can contain hundreds of genes!).

2.5 Network analysis

To extend the search of sets of genes beyond those pathways that are already known and encoded as distinct entities in databases, we would like to use a broader set of knowledge:

databases of interactions between genes, known as gene interaction networks. A broad range of databases contains gene interaction networks for *Homo Sapiens*. They include different types of data such as physical protein-protein interactions, other literature-curated interactions, co-expression interactions, yeast-to-hybrid interactions, inferred interactions from other species, etc. We would like to search for interacting genes that together contain a strong statistical signal of association with severe dengue. Interacting genes will be called *subnetworks* or *modules* in the following chapters.

2.5.1 Methods

Figure 2.5 gives an overview of the workflow used for the network analysis. We here use gene p-values generated by VEGAS that we map onto an available functional interaction network, use an algorithm to find subnetworks concentrating genes with low association p-values, and finally biologically interpret these subnetworks by testing for enrichment of certain Gene Ontology categories.

The network

Prior to network analysis, we need to choose a database of gene interactions. Some contain manually curated information only [Keshava Prasad et al., 2009], others use computational literature search and agglomeration of existing databases. There is, to my knowledge, no clear evaluation as to the quality/suitability of the different networks for different types of analyses. Since we wanted to include as many potential interactions as possible, we chose a network containing a broad variety of functional interactions. Each interaction is weighted according to a score designed to reflect the confidence in the existence of a given interaction, and based on the quality and quantity of available data.

I employed two different networks for my analyses. The first one was HumanNet [Lee et al., 2011], a functional interaction network spanning 476,398 scored functional interactions between 16,243 (87%) of validated human protein-coding genes. HumanNet uses only

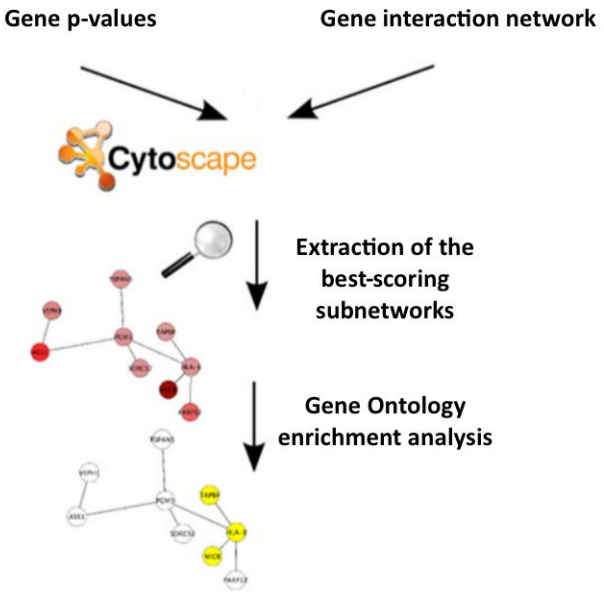


Figure 2.5: Workflow: Network analysis of GWAS results.

annotations supported by experimental evidence. Annotations are either inferred from a direct assay, inferred from mutant phenotype, inferred from a protein interaction, or inferred from a genetic interaction. To this network, we added 150 dengue-specific epistatic interactions that were available in our laboratory (unpublished work of Laura Grange). They had been detected using PLINK software [Purcell et al., 2007]. PLINK fast-epistasis mode uses a test based on a z-score for the difference in SNP-SNP association (odds ratio) between cases (dengue shock syndrome) and controls (non-disease samples). The top results had been confirmed by permutation analysis using MBMDR (10000 permutations/MaxT option) [Cattaert et al., 2011].

After HumanNet, I used STRING v9.1 [Franceschini et al., 2013], which appeared to be more frequently updated. This network contains, for *Homo Sapiens*, 4,319,956 interactions. This network aggregates data from several databases, literature text mining, predicted inter-

action based on homology, co-expression data, data from large-scale experiments, homology between similar species, and co-occurrence of protein domains.

Search tool

We are interested in bioinformatics tools that are able to take as input a network of interacting genes and SNP-level scores or gene-level scores, map these onto nodes of the network and search in the networks for subnetworks that aggregate high-scoring genes. A variety of search tools exist. Many of them had been initially designed for gene expression data. A review gives pointers to some of the methods [Jia and Zhao, 2013]. Table 2.3 shows those subnetwork prioritisation methods that I found to be suitable for SNP-level or gene-level p-value inputs.

Tool	Algorithm description
jActiveModules [Ideker et al., 2002]	Transforms node p-values into node z-scores, aggregates these scores using Stouffer's z-score method [Stouffer et al., 1949] and ensures that for each subnetwork size, scores follow a standard normal distribution. The user can choose between a greedy algorithm or simulated annealing to search for top-scoring subnetworks.
dmGWAS [Jia et al., 2011]	Same scoring function as jActiveModules, performs greedy search with two additional parameters. Parameter d controls the size of the space explored: each nodes needs to be within a distance d to any other node in the subnetwork. Parameter r controls whether a node should be added to the best solution: the node will be added if it improves the score of the best subnetwork by more than r times the current best score. It also computes p-values of results.
EW-dmGWAS [Wang et al., 2015b]	An adaptation of dmGWAS that includes edge weights into the subnetwork score.
PINBPA [Wang et al., 2015a]	A Cytoscape App [Shannon et al., 2003] that uses VEGAS output as input for jActiveModules, and computes significance using permutations.
PANOGA [Bakir-Gungor and Sezerman, 2011]	A pipeline suitable for SNP data on the basis of jActiveModules.
GXNA [Nacu et al., 2007]	Inspired by jActiveModules score. Attempts to correct for score dependencies in between connected nodes by introducing a parameter-dependent heuristic.
Bionet [Beisser et al., 2010]	Integer linear programming approach that is inspired by the Prize-Collecting Steiner Tree Problem. It optimises a scoring function based on estimating noise-to-signal ratio from node p-values. Then it develops an additive score, where positive values represent signal content and negative values represent background noise.
NIMMI [Akula et al., 2011]	First pre-computes a weight for each node based on Google PageRank algorithm, taking into account the numbers of neighbours and their neighbours using a dampening factor that, unlike Google PageRank, is scaled, and not constant. It then determines a combined subnetwork z-score as a sum of neighbouring scores weighted by their previously calculated weights. The available pre-computed weights of nodes have been calculated for the protein-protein interaction network BioGRID.
NetworkMiner [García-Alonso et al., 2012]	Takes as input a ranked list of genes. Finds subnetworks concentrating best-ranked genes using a gene partitioning approach.
SigMod [Liu et al., 2017]	Uses integer linear programming to optimise an objective function that is a weighted sum of gene scores and a weighted sum of edge scores, penalised by a fitted coefficient times the number of nodes in the subnetwork.

GWAStoNetwork [Hiersche et al., 2013]	Combines GWAS p-values p_A and p_B of connected genes A and B into an edge score (by default, $\log(p_A) \cdot \log(p_B)$). The graph partitioning algorithm then decomposes the entire network into subnetworks by concentrating high-weight edges within subnetworks and minimizing the total weights of between-subnetwork edges during the clustering process.

Table 2.3: Subnetwork search algorithms suitable for GWAS data

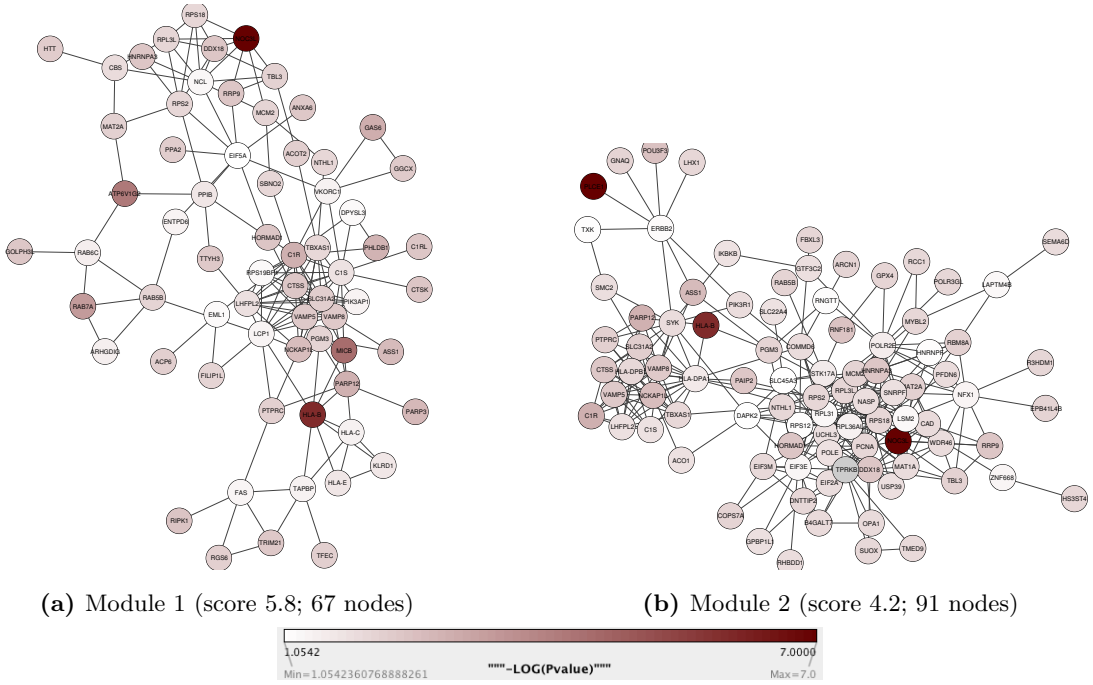
Among these algorithms, I chose, for a first evaluation, the one that I found most widely used and cited: jActiveModules [Ideker et al., 2002]. The obtained sets of genes being too large to be analysed one by one, I used the gene ontology (GO) enrichment tool BINGO [Maere et al., 2005] to perform a hypergeometric test. BINGO determines which Gene ontology (GO) terms are significantly overrepresented in the set of genes of interest. Gene ontology is a bioinformatics community resource to annotate genes using predefined terms, enabling genes to be directly grouped by these terms.

2.5.2 Results

When run with the entire network, jActiveModules did not terminate within 48 hours. I then reduced the input network to the interactions between those top 10% genes that had the best p-values. Results using STRING and HumanNet networks, along with Gene ontology (GO) enrichment of the resulting genes are displayed in Figures 2.6, 2.7, 2.8, and 2.9.

The best-scoring subnetworks tend to include some of the genes with the lowest p-values. When using the GO enrichment tool BINGO [Maere et al., 2005], on HumanNet network, the MHC complex genes again appear as an enriched category; "Antigen processing and presentation of peptide antigen via MHC class I" has a multiple-testing corrected enrichment p-value of 3.10^{-4}. Genes from the network that fall within this category are: TAPBP, HLA-B, HLA-C, and HLA-E.

The complement activation classical pathway is also enriched with a p-value of 0.02. The complement system is a part of the immune system that complements the ability of antibodies and phagocytic cells to clear microbes and damaged cells from an organism, by promoting inflammation via cytokines, and attacking the pathogen's plasma membrane. It



(a) Module 1 (score 5.8; 67 nodes)

(b) Module 2 (score 4.2; 91 nodes)

(c) Gene p-value color scale

Figure 2.6: Top-scoring subnetworks using jActiveModules tool and HumanNet network.

is part of the innate immune system, but it can be recruited and brought into action by the adaptive immune system (*cf.* Chapter 1). Genes in the subnetwork belonging to this category are: C1RL, C1R, C1S.

When using the STRING interaction network, enriched categories are very different from the ones that we had previously obtained with HumanNet; only groups related to kidney development are significantly enriched with a corrected p-value of 0.002. Genes from our subnetwork belonging to this category are: FOXC2, PLCE1, ASS1, POU3F3, PYGO1, and AGTR1. AGTR1, or angiotensin II is a potent vasopressor hormone (i.e., it stimulates contraction of the muscular tissue of the capillaries and arteries) and a primary regulator of aldosterone secretion. It is an important effector controlling blood pressure and volume in the cardiovascular system. Blood pressure and volume are key parameters in the most severe form of dengue, dengue shock syndrome: most severe patients have heart failure that can occur because of insufficient blood pressure.

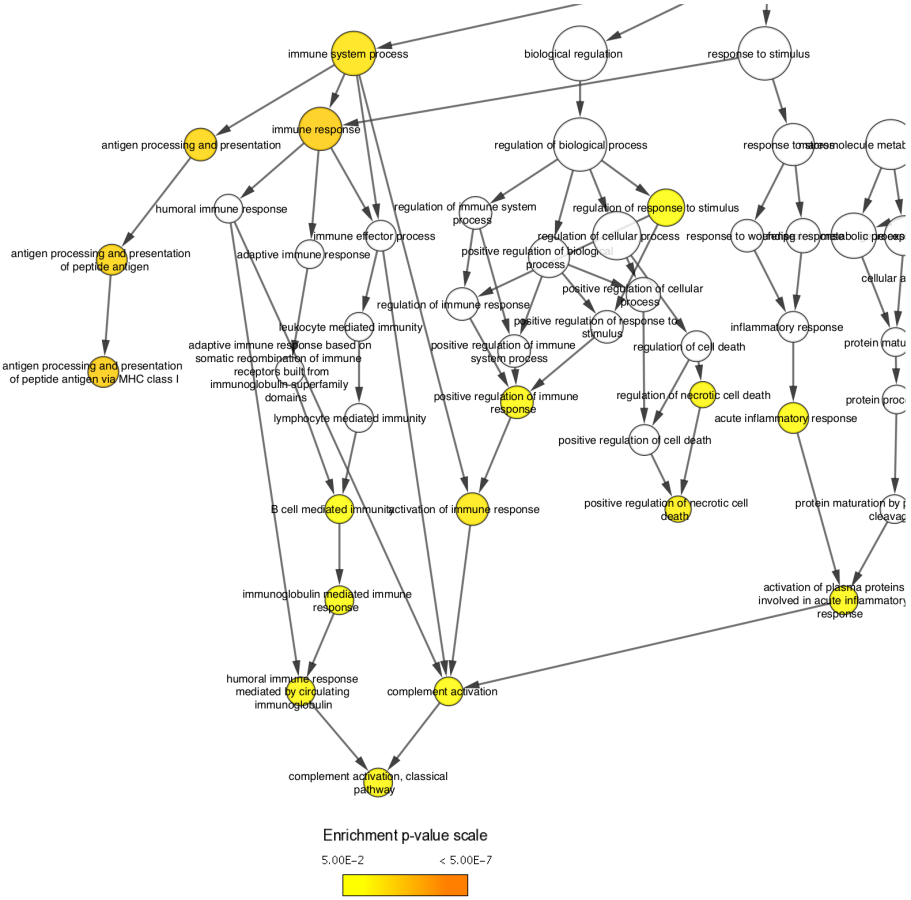
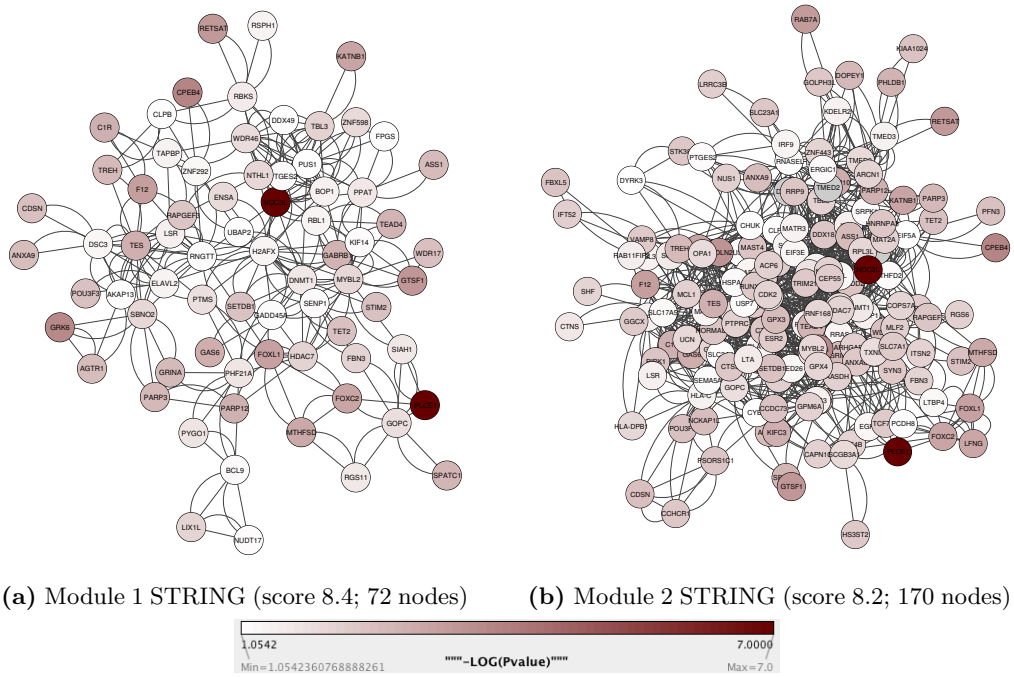


Figure 2.7: Gene ontology enrichment of the best subnetwork in immune processes.

2.5.3 Discussion

Network analysis using HumanNet confirms that immune activation may play a role in severe dengue susceptibility. Additionally, the resulting subnetwork is enriched in genes from the complement activation classical pathway.

STRING network analysis using the same gene p-values generates a very different result, with much higher network scores, but enriched in a completely different category: kidney development. This enriched category points in a similar direction as PLCE1 (*cf.* the discussion in [Khor et al., 2011]). Mutations within PLCE1 are associated with nephrotic syndrome [Hinkes et al., 2006], a kidney disorder that, when severe, leads to reduced vascu-



(a) Module 1 STRING (score 8.4; 72 nodes) **(b)** Module 2 STRING (score 8.2; 170 nodes)

(c) Gene color scale

Figure 2.8: Top-scoring subnetworks using jActiveModules tool and STRING network.

lar oncotic pressure and edema. Oncotic pressure is a form of osmotic pressure exerted by proteins, notably albumin, in the plasma of a blood vessel that usually tends to pull water into the circulatory system, suggestive of a link between low quantities of fluid in blood and PLCE1. Moreover, another gene in the same gene ontology category, AGTR1, is a potent vasopressor hormone, and an important effector controlling blood pressure and volume in the cardiovascular system. These elements together strengthen the hypothesis that genetic predisposition to severe dengue is associated with genes regulating blood pressure and maintaining normal vascular endothelial cell barrier function in this dataset.

How is this related to clinical manifestations? When a patient has an infection in a given place in the body, the inflammation signal increases blood vessel permeability; plasma then gets more easily to the origin of inflammation, carrying with it clotting factors to stop the bleeding and spread of infection, antibodies to fight infection, nutrients to feed the tissue cells, and proteins that attract phagocytes [Luft, 1965]. When this process happens locally, it helps the body to heal faster, and the loss of fluid in blood is small. In dengue, the

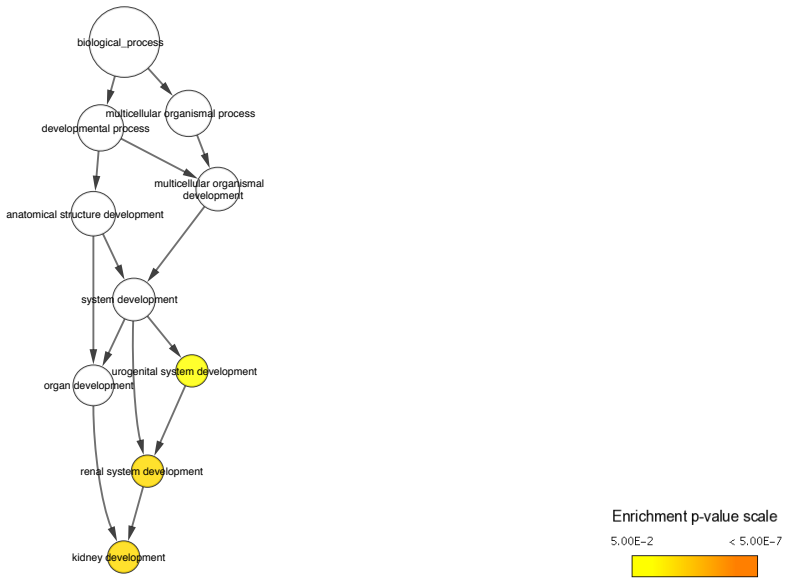


Figure 2.9: GO enrichment of subnetwork 1, STRING network.

virus spreads to the entire organism, thus generates a systemic inflammation and capillary permeability. If badly managed by the body, this may result in too much plasma leaking out of blood vessels, and, together with an inability to compensate for the lost blood volume, the possibility of heart failure (clinical shock). Plasma leakage and shock are characteristic symptoms of severe dengue.

From a methodological point of view, the difference in the results from two different networks is intriguing. In both cases, we only used the network of 10% genes with the lowest p-values. Therefore, even though these networks tend to be highly connected, the connectivity between genes plays a major role for the result. Moreover, the method we employed does not provide a measure of statistical significance, which leaves open the possibility that our results may not be statistically significant.

Additionally, the best subnetwork score obtained form STRING interaction network is 45% higher than the one from HumanNet. However, the HumanNet best subnetwork is enriched in more Gene Ontology categories that point towards the same immune-related process. The hypothesis that this process is involved in severe dengue thus appears to be more robust.

A possible explanation may be that there are fewer genes annotated in best subnetworks obtained using STRING results. Other causes for the difference between results obtained from different networks may lie in certain undocumented properties of the jActiveModules tool. For instance, in its default configuration, jActiveModules runs with an activated option correction for subnetwork size that is not documented in any publication or website, to my knowledge. The name of this option suggests that it represents an attempt to correct for an issue with jActiveModules, namely the empirical observation that jActiveModules has a size bias, i.e., that it tends to return very large subnetworks as results. In the following chapter, I present an explanation and an analysis of this phenomenon from a theoretical point of view.

The results are also difficult to interpret beyond gene set enrichment: gene-by-gene exploration of my resulting subnetworks of 67 genes or more is time-consuming, and it is still unclear whether it would lead somewhere given the information losses/ad-hoc choices at different stages of the analysis pipeline: SNP-to-gene mapping (as discussed above), reduction of the analysis to 10% top-scoring genes, thus removing connections in between genes, etc.

2.6 Conclusions

Results from GWAS, gene-level p-value aggregation using VEGAS, and network analysis using jActiveModules, all suggest that immune activation plays a key role in dengue susceptibility in this dataset, as well as kidney-related genes, implicated in regulating blood pressure, and in maintaining normal vascular endothelial cell barrier function.

In VEGAS results, HLA-B, MICA, HCP5, NOC3L, and PLCE1 are significantly associated with severe dengue. The first three genes are part of the MHC (major histocompatibility complex), whose function is to recognise pathogens and display them on the cell surface, so that the appropriate T cells can recognise them. This result adds information to the initial GWAS result, in which only one gene from this complex was found to be associated

to severe dengue. PLCE1, whose disfunctions had been reported to be related to a kidney disease, had already been significantly associated with severe dengue by the GWAS. This result remains when mapping SNPs to genes.

On the one hand, the subnetwork obtained using the HumanNet network expands on gene-level VEGAS results by finding subnetworks enriched in antigen processing and presentation of peptide antigen via MHC class I. On the other hand, this result is also enriched in genes from the complement activation classical pathway. This pathway complements the same ability of the human immune system to recognise pathogens, create inflammation via cytokine release and pathogen removal. STRING network analysis expands on the kidney-related GWAS result (the PLCE1 gene) by detecting enrichment in the in "kidney development" category. Genes falling within this category include genes not only related to blood volume, but also stimulation of contraction of the muscular tissue of capillaries and arteries.

The methodology applied here required several ad-hoc choices or parameters during the mapping of SNPs to genes, gene p-value computation, choice of the input network, and subnetwork search algorithm. The following chapter will focus on issues related to the subnetwork search algorithm, and Chapter 4 will present an alternative search tool that addresses some of the problems in jActiveModules and emphasizes interpretability.

Chapter 3

Towards an unbiased score function for identifying network modules

Contents

This chapter has been submitted as an article in April 2017 and is now under review. It discusses the bias of the scoring of some subnetwork search methods such as jActiveModules.

3.1 Chapter summary

Biological processes often manifest themselves as coordinated changes across several interacting molecules in high-dimensional data. Such data is therefore often visualized and analyzed in the context of interaction networks. In these networks, subnetworks that may correspond to correlated change can then be identified through computational search. According to several reports, one of the first and frequently used subnetwork scores for this problem, introduced in the jActiveModules software, has a strong tendency to lead to large subnetworks. Follow-up versions of the method have dealt with this issue only by introducing *ad hoc* corrections whose efficacy remains limited.

Here, we show that the size bias is not only an empirical phenomenon for specific datasets, but a statistical property of the underlying score function. Based on this, we present a new score function that removes the size bias. A sampling approach to computing the new score function is computationally hard, but we present evidence that the score can be approximated using extreme value functions.

3.2 Introduction

The organisation of cells is thought to be inherently modular [Alon, 2003, Hartwell et al., 1999]. When studying large-scale datasets, a common approach to identify those modules relevant to a question of interest starts with experimental or other gene-level scores that indicate some level of involvement of genes in a biological question, and to then identify modules with aggregate scores that are higher than expected by chance.

In such an approach, modules can either consist of predefined gene sets, such as pathways [Khatri et al., 2012], or connected subnetworks of a network of interacting genes [Mitra et al., 2013]. Predefined gene sets have the advantage of being easier to analyse and interpret, but are obviously limited by existing knowledge. Functional interaction networks represent information on pairs of genes known to interact—directly or indirectly—in the same biological context. The nodes of such networks typically represent macromolecules, such as proteins. Edges can represent hypothetical or verified physical associations, such as protein-protein, protein-DNA, metabolic pathways, DNA-DNA interactions, or functional associations, such as epistasis, synthetic lethality, correlated expression, or correlated biochemical activities [Szklarczyk et al., 2014, Keshava Prasad et al., 2009, Lee et al., 2011].

Modules are typically identified as subnetworks with high aggregate gene-level scores. Aggregation is typically performed using a *normalised* score function whose distribution is identical for all subnetworks sizes in a null model.

Many algorithms are based on the score defined by jActiveModules [Ideker et al., 2002], including PANOGA [Bakir-Gungor and Sezerman, 2011], dmGWAS [Jia et al., 2011], EW-dmGWAS [Wang et al., 2015b], PINBPA [Wang et al., 2015a], GXNA [Nacu et al., 2007], and PinnacleZ [Chuang et al., 2007]. These methods are widely applied in the current literature [Sharma et al., 2013, Olex et al., 2014, Smith et al., 2014, Pérez-Palma et al., 2016, Jin et al., 2008, Chuang et al., 2007, Dao et al., 2011, Liu et al., 2007, Qiu et al., 2010, Hormozdiari et al., 2015], even though the above approaches have been reported to consistently result in subnetworks that are large, and therefore difficult to interpret biologically [Nacu et al., 2007, Rajagopalan and Agarwal, 2005, Batra et al., 2017]. Some versions of the approach have dealt with this issue by introducing heuristic corrections designed to remove the tendency towards large subnetworks [Nacu et al., 2007, Rajagopalan and Agarwal, 2005, Liu et al., 2017]. A recent evaluation of several algorithms revealed that the efficacy of these corrections remains limited [Batra et al., 2017]. Other methods avoid dealing with the issue by allowing users to limit the size of the returned module [Jia et al., 2011, Wang et al., 2015b, Wang et al., 2015a, Nacu et al., 2007, Chuang et al., 2007, Beisser

et al., 2010], which is problematic, as users typically do not have prior information about suitable settings of this parameter.

Here, we find that this tendency is not just a capricious property of selected datasets, but that a fundamental size bias is built into the score function itself. This leads us to define a new score function that is free of size bias. We show that, even though the practical approximation of the background distribution by sampling is computationally hard, extreme value distributions may provide good models. In the light of these results, we provide our view of the currently best options for avoiding the size bias.

3.3 Materials and Methods

3.3.1 The subnetwork identification problem

Most of the above-mentioned module identification methods are motivated as a maximisation problem over a set of (connected) subnetworks of a graph. In its basic form, its three inputs can therefore be described as follows.

1. A graph G, corresponding to the functional interaction network, in which the nodes $V = (v_1, ..., v_N)$ correspond to molecules. By $\mathbf{A}(G)$ we denote the sets $A \subseteq V$ that induce connected *subnetworks* in G. By $\mathbf{A}_k(G)$ we denote only those sets of size $|A| = k$, which we will also call *k-subnetworks*.

2. A set of P-values $(p_1, ..., p_N)$ that correspond to the statistical significance of observations associated with the N molecules.

3. A score function $s(A) : \mathbf{A}(G) \longrightarrow \mathbb{R}$ that assigns a score to each connected subnetwork.

A solution to the subnetwork identification problem corresponds to a subnetwork A that maximises the score $s(A)$ over $\mathbf{A}(G)$.

3.3.2 jActiveModules score function

The jActiveModules method [Ideker et al., 2002] was one of the first published subnetwork identification methods. Given an input graph G and P-values $(p_1, ..., p_N)$, a first aggregate score $z(A)$ for a k-subnetwork $A \in \mathbf{A}_k(G)$ is defined using Stouffer's Z-score method [Stouffer et al., 1949]:

$$z(A) = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i,$$

where $z_i = \phi^{-1}(1 - p_i)$, and ϕ^{-1} is the inverse normal cumulative distribution function (CDF). The jActiveModules score $s(A)$ is then obtained as

$$s(A) = \frac{z_A - \mu_k}{\sigma_k},$$

where μ_k and σ_k are sampling estimates of mean and standard deviation of scores z_A over all k-node sets $A \subseteq V$. Ideker et al. [Ideker et al., 2002] evaluated the resulting score against a distribution of empirically obtained scores under random permutations of $(p_1, ..., p_N)$, corresponding to a null hypothesis of a random assignment of input gene-level scores to the nodes of the network.

3.4 Definitions

To discuss the key subnetwork score properties that are at the origin of the size bias, we introduce the following notations.

3.4.1 Subnetwork scores S_k, S_k^*

By S_k we denote a random variable that describes the occurrence of *k-subnetwork scores*, with CDF $F(x) = P(s(A) \leq x \,|\, A \in \mathbf{A}_k(G))$. Similarly, we denote by S_k^* the *maximal* k-

subnetwork scores with CDF $F(x) = P(\max_{A \in \mathbf{A}_k(G)} s(A) \leq x)$. Below, we will discuss the distributions of S_k and S_k^* under the null hypothesis.

3.4.2 Score normalisation

Per construction of the jActiveModules score function, and under a sufficient amount of sampling to determine μ_k and σ_k, S_k follows a standard normal distribution: $S_k \sim \mathcal{N}(0,1)$[Ideker et al., 2002]. Whenever, as here, the distribution of S_k is independent of k, we will call the underlying score s *normalised*. As we will show below, the size bias of the jActiveModules approach is rooted in the fact that the underlying score is normalised.

3.5 Empirical studies of small subnetworks and their scores

We show that, under a normalised score, small subnetworks can be significantly high-scoring in their size class, but still low-scoring when compared to scores that occur by chance in larger networks, thus explaining the above-mentioned size bias, *i.e.,* the tendency of jActiveModules and related methods to return large subnetworks.

To empirically explore the properties of the jActiveModules score function, we generated a sample network with 50 nodes from STRING interaction network [Szklarczyk et al., 2014], which we denote by G_{50}, by first initialising a graph G_{current} with a randomly chosen node from the STRING network. Then we iteratively extended G_{current} with a randomly chosen neighbour, until $|G_{\text{current}}| = 50$.

3.5.1 For small values of k, the number $|\mathbf{A}_k(G)|$ of k-subnetworks increases strongly with k

By definition, the null distribution of a normalised score over all k-subnetworks is identical for all values of k. What normalisation does not take into account is the fact that the

number $|\mathbf{A}_k(G)|$ of k-subnetworks depends on k.

We now explore this effect for different graphs G. In a fully connected graph G, each k-subset $A \subseteq V$ forms a k-subnetwork. Here, $|\mathbf{A}_k(G)| = \binom{N}{k}$, which strongly increases with increasing small k.

Figure 3.1 shows that, also for our sample network $G = G_{50}$, $|\mathbf{A}_k(G)|$ strongly increases with k for small k.

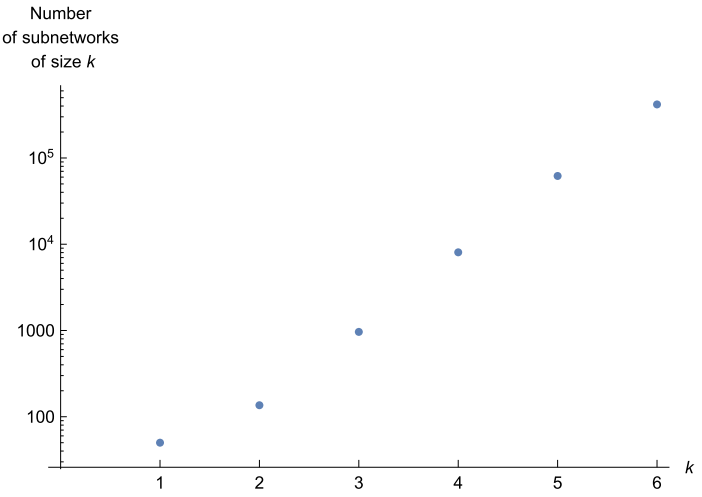


Figure 3.1: Numbers $|\mathbf{A}_k(G)|$ of small subnetworks in G_{50} (a network of 50 nodes) as a function of their size k

Finally, the STRING [Szklarczyk et al., 2014] network G with 250000 highest-scoring edges has $|\mathbf{A}_3(G)| = 20676496$ 3-subnetworks, and $|\mathbf{A}_4(G)| = 201895916$ 4-subnetworks. The number of 5-subnetworks was higher yet; we were not able to calculate $|\mathbf{A}_5(G)|$ in a reasonable amount of time.

3.5.2 Maximum scores S_k^* increase strongly with k under the null hypothesis

We now explore the behaviour of the maximum k-subnetwork score S_k^* under the null hypothesis, with increasing k, for small values of k. As $|\mathbf{A}_k(G)|$ tends to increase strongly

with small k (Section 3.5.1), and the distribution of jActiveModules scores S_k is independent of k (cf. Section 3.4.2), one may expect S_k^* to strongly increase with k. Figure 3.2 illustrates this effect in the case of i.i.d. samples.

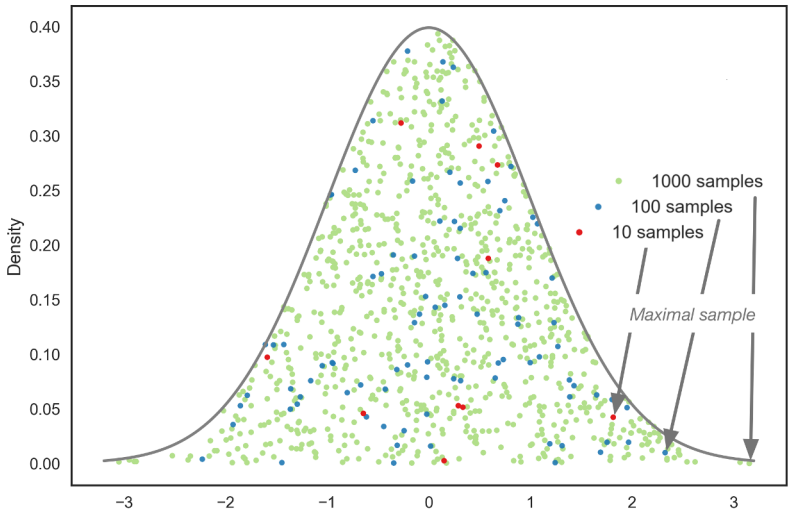


Figure 3.2: Sample maxima from i.i.d. samples are likely to increase with sample size.

Subnetwork scores S_k are not independent, as subnetworks in $\mathbf{A}_k(G)$ are overlapping. To explore whether the same effect as in the independent case can still be observed, we computed scores S_k^* in our sample network $G = G_{50}$ for 100000 random instantiations of $(p_1, ..., p_{50})$. Figure 3.3 shows the resulting empirical distributions of S_k^*, for some small values of k, with a clear increase of S_k^* with increasing k.

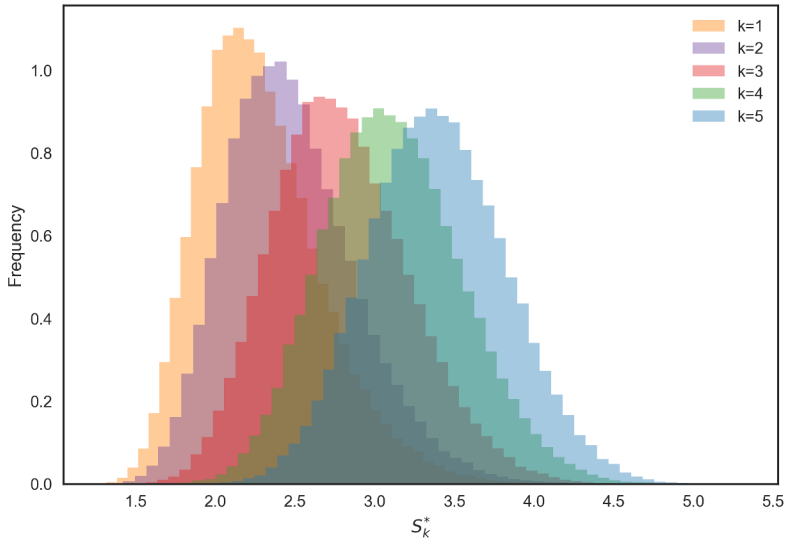


Figure 3.3: Empirical distributions of jActiveModules maximum subnetwork scores S_k^* in the graph G_{50} for small values of k under the null hypothesis

We note in passing that, for large values of k, the number $|\mathbf{A}_k(G)|$ of connected subnetworks *decreases* with k (in particular, $|\mathbf{A}_N(G)| = 1$ for connected graphs G). Accordingly, one may expect decreasing maximum scores S_k^* when k becomes close enough to N. Our empirical evaluation, shown in the Appendix of this Chapter (Figure 3.6), is consistent with this idea: On our sample graph G_{50}, jActiveModules scores S_k^* *decrease* for $k = 46, 47, 48$.

3.5.3 Maximum scores S_k^* may follow an extreme value distribution under the null hypothesis

Maxima of independent identically distributed (i.i.d) scores follow an extreme value distribution [Coles, 2001]. Subnetwork scores are indeed identically distributed: they follow a standard normal distribution (Figure 3.7). However, due to the overlap between subnetworks, subnetwork scores S_k are not independent. Nevertheless, most pairs of *small* subnetworks of a larger network do not overlap, and their dependency structure is therefore

local.

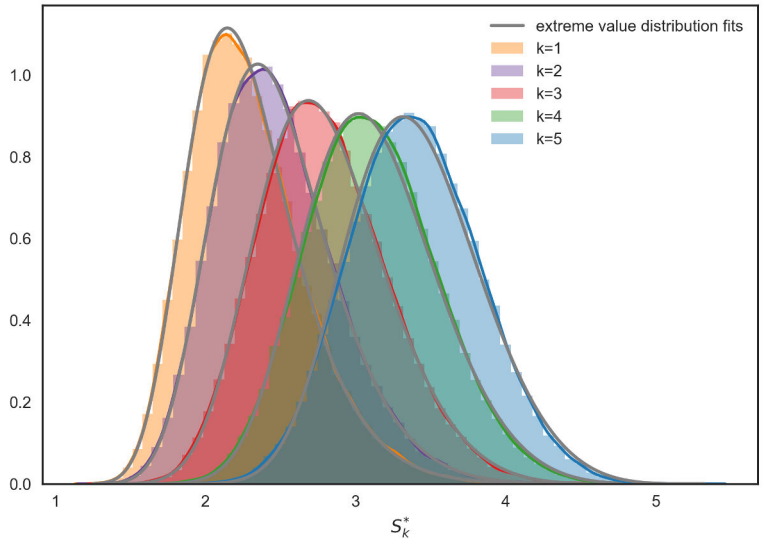
Extreme value distributions are used in other cases when dependency structure is local. They have been been proved to accurately approximate certain sequences of random variables whose high scores (*block maxima*) have a local dependency structure [Coles, 2001]. In sequence alignment, high-scoring alignments tend to overlap locally, and Karlin and Altschul [Karlin and Altschul, 1990] demonstrated that the null distribution of local similarity scores can be approximated by an extreme value distribution. There, a weighting parameter K explicitly accounts for the non-independence of the positions of high-scoring matches. K is specific to the search database, and its estimation is computationally intensive.

Figure 3.4 shows that generalised extreme value distributions also fit empirically observed distributions S_k^* quite well in the sample network G_{50} with its fit parameters (Probability plots in Section 3.8.3). The fit can be observed to be good for smaller values of k, and to deteriorate with increasing k, concomitant with the loss of locality in the subnetwork dependency structure.

3.6 Discussion

3.6.1 The jActiveModules score and other normalised scores are biased towards larger subnetworks

Our empirical study of maximal subnetwork scores suggests that maximum scores S_k^* strongly increase under the null hypothesis when k is small (Section 3.5.2, Figure 3.3). This implies that certain non-significant subnetworks of larger size are systematically scored higher than other, smaller, subnetworks that have a significantly high score relative to their size. Figure 3.5 illustrates this effect: a score that is unlikely to be observed by chance in a 3-subnetwork is much more likely to be observed by chance in a 5-subnetwork. Even though we were not able to explicitly calculate S_k^* for $k > 5$, we deem it likely that, larger



(a)

k	μ_k	σ_k	ξ_k
1	2.1	0.33	0.08
2	2.3	0.36	0.12
3	2.6	0.40	0.16
4	2.9	0.41	0.18
5	3.2	0.42	0.20

(b)

Figure 3.4: (a) Fits of generalised extreme value distributions $F(x; \mu_k, \sigma_k, \xi_k)$ to empirical distributions of S_k^* . Colored lines represent the smoothed versions of the histograms, whereas the grey lines are fits from the family of extreme value distributions, and (b) the parameters of the fits.

k-subnetworks (with, say, $k > 7$) with even better scores are almost certain to exist in random data. As many methods do not provide an assessment of the statistical significance of the reported subnetworks, these methods not only prefer spurious larger subnetworks over—potentially biologically relevant—smaller ones, but also fail to provide their users with an indication that the reported networks are indistinguishable from chance observations.

3.6.2 An unbiased score function $\tilde{s}$

It is straightforward to remove the bias of a (normalised or unnormalised) score $s(A)$ by calibrating it relative to its size-specific null distribution. For a k-subnetwork A, one can define

$$\tilde{s}_k(A) = P(S_k^* \leq s_A).$$

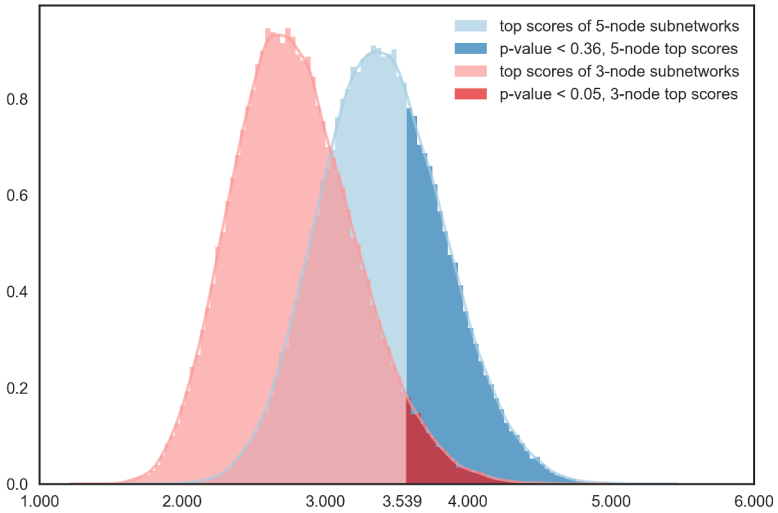


Figure 3.5: Scenario illustrating the bias of normalised scores towards larger subnetworks. Distributions shown are jActiveModules null distributions S_3^* and S_5^* for the sample network G_{50}. Under the null hypothesis, a score of 3.539 that is unlikely to occur for a 3-subnetwork ($P(S_3^* \geq 3.539) \approx 0.05$) is much more likely to occur for a 5-subnetwork ($P(S_5^* \geq 3.539) \approx 0.36$).

For each k, the resulting maximum scores $\tilde{S}_k^*$ are then approximately uniformly distributed on $[0, 1]$, i.e., $P(\tilde{S}_k^* \leq x) \approx x$. Note that the uniform distribution is only approximate, as $\tilde{S}_k^*$ is a discrete distribution.

3.6.3 Computing the unbiased score $\tilde{s}$ by sampling is computationally hard, but it may be possible to approximate $\tilde{s}$ by an extreme value distribution

Computing the above score function $\tilde{s}$ is not straightforward. In principle, $\tilde{s}(A)$ could be approximated by sampling from S_k^*, but each sample requires the computation of a maximum of $s(A)$ over all subnetworks A in a network whose gene-level scores have been instantiated with P-values — a problem that has been shown to be NP-hard even in a simplified form [Ideker et al., 2002]. Approaches to solve this problem nonetheless exist [Dittrich et al.,

2008, Liu et al., 2017], but under the reported running times in the range of minutes to hours for a single sample from S_k^*, sampling still remains very time-consuming.

The locality of the dependency structure among small subnetworks and our empirical results from Section 3.5.3 suggest that S_k^* can possibly be approximated by an extreme value distribution. However, it is not obvious how the parameters of this distribution can be estimated practically without recourse to sampling, which, as discussed above, is difficult.

3.6.4 Current options to avoid size bias

In the absence of practical solutions to compute the unbiased subnetwork score $\tilde{s}$, what are the current practical options for scoring and detecting subnetwork aggregates of statistical signals?

One possibility is to use one of the approaches that find highest-scoring subnetworks of a fixed, or limited, subnetwork size k [Backes et al., 2012, Jia et al., 2011, Wang et al., 2015b, Wang et al., 2015a, Nacu et al., 2007, Chuang et al., 2007, Beisser et al., 2010], and to compare them on the basis of their biological interpretation. Since only small networks tend to be biologically interpretable, only small k would have to be tested. As adding a few neighbours to a statistically significant subnetwork can be expected to preserve significance, not all values of k would need to be tested. While this approach has obvious shortcomings (solutions for different values of k need to be compared, multiple statistical tests, sometimes unclear biological interpretation), each computation by itself would only compare subnetworks of same size, and thus avoid size bias.

There are other, non-statistical (e.g., algorithmic/physical) principles for identifying aggregates of signals in networks [West et al., 2013, Alcaraz et al., 2014]. The lack of clear mathematical relationships between inputs and outputs, and the lack of options to assess statistical significance may make it difficult to evaluate these approaches, and their applicability to any given biological scenario. We have developed an approach that preserves mathematical clarity and statistical tools, and obtains computational tractability through a

restriction to a simplified subnetwork model. This approach, LEAN is developed in Chapter 4 and published in [Gwinner et al., 2016].

3.7 Conclusions

The identification of functional subnetworks of strongest aggregate statistical signals in networks is an important approach to analyse biological genome-scale datasets. An array of different computational methods and software is in practical use, but many are plagued in practice by a recognised strong tendency towards large subnetworks that *ad hoc* adjustments have not been able to remedy.

Here, we present a first direct analysis of the origins of this phenomenon that reveals a strong statistical size bias in a frequently used score function. By normalisation against size-specific null distributions, we derive a new, unbiased, score. This score function is computationally hard, and we outline our view of currently best other practical options to avoid size bias. Finally, we hope that our evidence, that the unbiased score function can be approximated using extreme value functions, can motivate further theoretical developments towards the unbiased identification of modules in networks.

3.8 Appendices to this chapter.

3.8.1 For large values of k, maximal subnetwork scores *decrease*

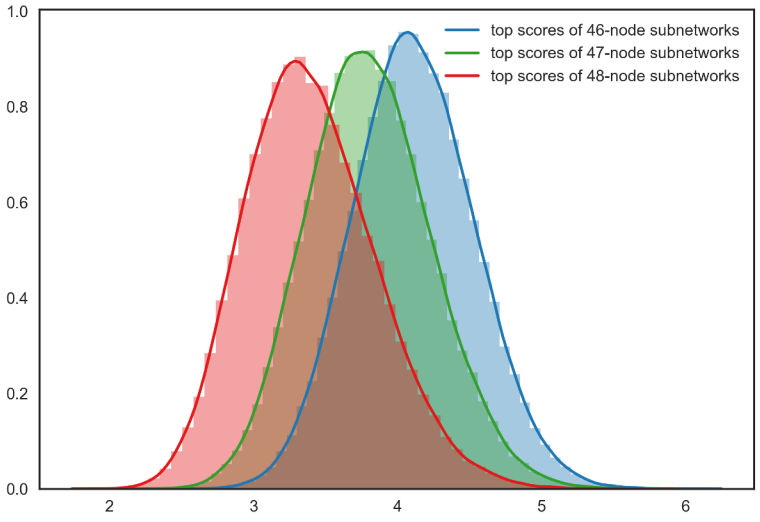


Figure 3.6: Distributions of maximum subnetwork scores S_k^* for large values of k under the null hypothesis.

3.8.2 Approximate normality of subnetworks scores S_k

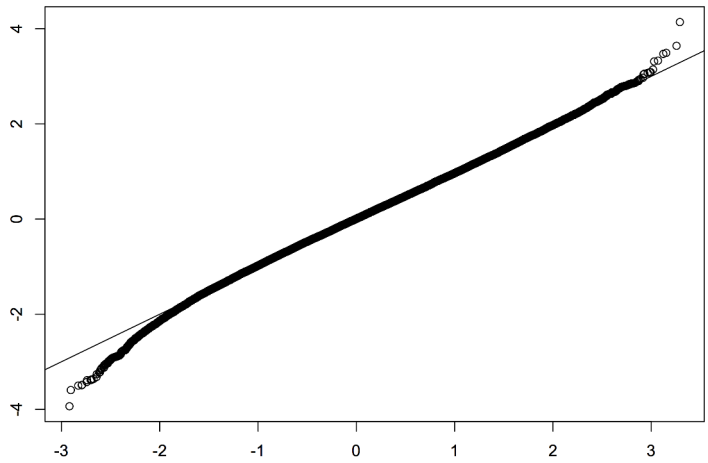


Figure 3.7: Quantile-quantile plot between standard normal distribution and jActiveModules scores S_5 for the sample graph G_{50} under the null hypothesis. Other scores S_k have similar quantile-quantile plots (not shown).

3.8.3 Quality of extreme value distribution fits for maximal subnetwork scores S_k^*

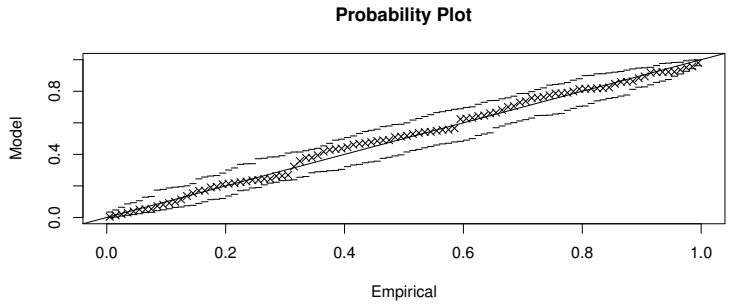


Figure 3.1: Probability plot for the extreme value model fit to maximal scores of subnetworks of size 1, S_1^*, in G_{50}.

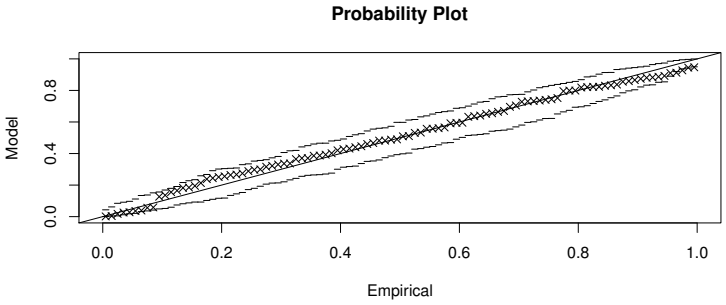


Figure 3.2: Probability plot for the extreme value model fit to S_2^*.

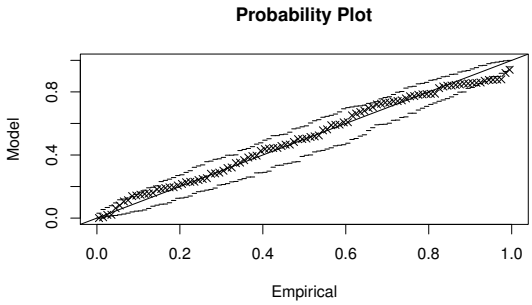


Figure 3.3: Probability plot for the extreme value model fit to S_3^*.

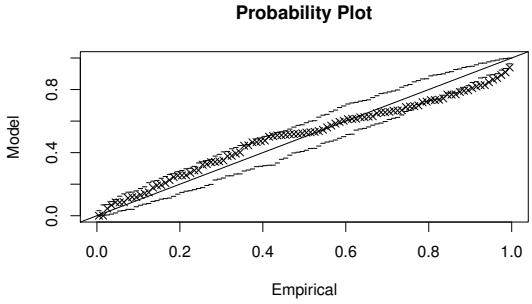


Figure 3.4: Probability plot for the extreme value model fit to S_4^*.

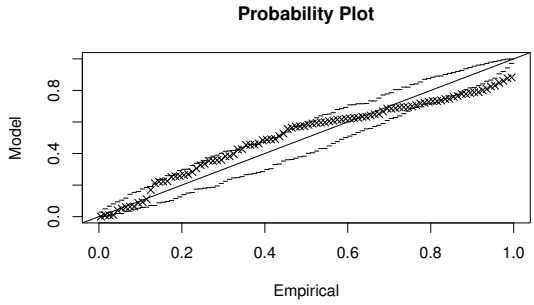


Figure 3.5: Probability plot for the extreme value model fit to S_5^*.

Chapter 4

The LEAN algorithm and its application to dengue data

Contents

In the previous chapter, I discussed the tendency of popular algorithms for subnetwork identification to return large subnetworks that are hard to interpret, while requiring the user to set many parameters with little obvious guidance. In this chapter, I describe the Local Enrichment Analysis (LEAN) method, which I co-designed, and that attempts to avoid these issues. LEAN has been implemented in an R package, first been applied on biological data by Frederik Gwinner. The first part of this chapter is an adaptation of the methods part of our article [Gwinner et al., 2016] that explains the algorithm. In the second part, I apply LEAN to dengue transcriptomic data.

4.1 The LEAN algorithm

4.1.1 Main idea: The local subnetwork model

We introduce here a novel network-based analysis approach integrating genome-wide measures of statistical significance (p-values) with large-scale interaction networks. It is based on a local subnetwork model that assumes that higher-order biological activity can be detected by aggregating signals from a single gene and its direct network neighbors (*cf.* Figure 4.1). The local subnetwork model is much simpler than the common (unconstrained) subnetwork model, in terms of computational complexity, and the assessment of statistical significance. While the number of subnetworks is typically exponential in the number of genes, networks contain only a single local subnetwork per gene. The identification of optimal subnetworks is computationally NP-hard [Ideker et al., 2002], whereas optimal local subnetworks can be identified in polynomial time by examining all genes and their neighborhoods in turn. The relatively low number of local subnetworks also allows the straightforward calculation of empirical p-values, while, for many subnetwork-based analysis methods, no efficient algorithms are known to compute statistical significance.

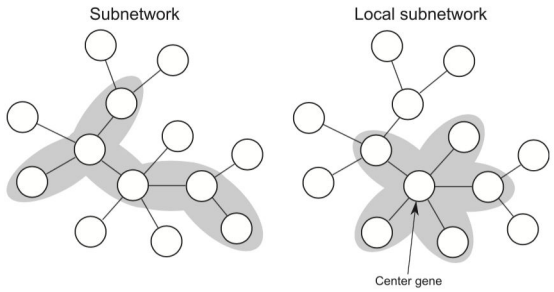


Figure 4.1: Subnetwork and local subnetwork pathway models. Local subnetworks are specific subnetworks that consist of a center gene and its direct network neighbors.

4.1.2 Local enrichment analysis

LEAN is based on two ingredients: A list of measures of statistical significance (p-values) for some or all genes and an interaction network. In many applications, p-values originate from a statistical test for differential expression, such as a t-test. While the approach is readily applicable to other types of datasets, we will describe it using the example of its application to the results of a differential expression analysis (input p-values). Analysis is performed using the given interaction network restricted to genes for which an input p-value has been calculated based on transcriptomic data. A local subnetwork A_g consists of a subset of genes formed from a center gene g and its directly interacting partners in the given network. Candidate subnetworks are all local subnetworks A_g.

4.1.3 LEAN p-values

For each candidate subnetwork A_g of size m, LEAN aims to evaluate whether for any $k \in \{1, \ldots, m\}$, the k genes of A_g with the best scores (e.g., lowest p-values) are statistically enriched for extreme scores (low p-values). To this end, an *unnormalized enrichment score* ES_g is computed on the basis of the sorted sequence of gene scores $p_1 \leq ... \leq p_k \leq ... \leq p_m$ of genes in A_g. To compute ES_g, for each position $k = 1, ..., m$ in the sorted subnetwork p-value list, we first calculate the probability $\tilde{p}_g^{(k)}$ that, under the null hypothesis of input

p-values being independent and identically distributed (i.i.d), and being sampled from a uniform distribution, at least k of the p_i are lower or equal to p_k using the cumulative distribution function of the binomial distribution:

$$\tilde{p}_g^{(k)} = \sum_{i=k}^{m} p_k^i (1 - p_k)^{m-i}. \tag{4.1}$$

We designate the position in the ordered subnetwork p-value list of A_g at which the minimum $\tilde{p}_g^{(k)}$ is achieved by $k^* = \arg\min \tilde{p}_g^{(k)}$. The *unnormalized enrichment score* ES_g is then defined as:

$$ES_g = \log_{10}(\tilde{p}_g^{(k^*)}). \tag{4.2}$$

To correct for biases due to subnetwork size, and to evaluate statistical significance, the enrichment p-value p_g^* is computed by comparing ES_g to a background distribution of ES_{BG} values obtained on random gene sets of the same size as A_g:

$$p_g^* = prob(ES_{BG} \geq ES_g). \tag{4.3}$$

To determine the background distribution of ES_{BG} values, p_g^* is empirically estimated using 10 000 (a user-configurable parameter) random samples of size m from the set of input p-values. To correct for the number of local subnetworks being tested, a Benjamini-Hochberg multiple testing correction is applied to the p-values of all candidate subnetworks. These multiple testing corrected p-values are further called the LEAN p-values. For each candidate subnetwork with a significant LEAN p-value, the LEAN implementation returns its central g gene along with the above mentioned intermediate scores and additional information on the candidate subnetwork. Figure 4.2 provides an example for the calculation of p_g^* for a candidate subnetwork of size $m = 7$.

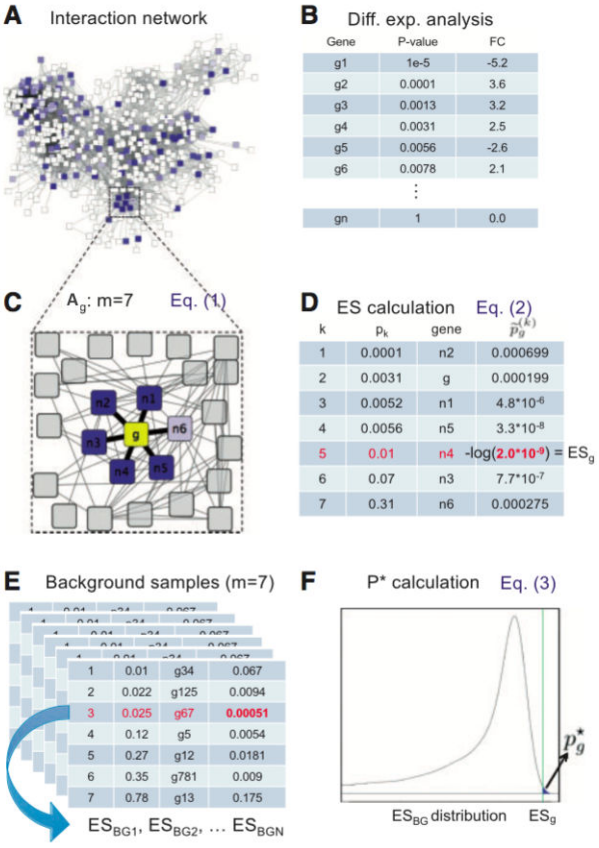


Figure 4.2: Summary of LEAN. Inputs are (A) an interaction network and (B) an input p-value for each gene in the network. For any gene g, the genes in its direct neighborhood along with their individual input p-values are then extracted from the network (C). The p-values within the neighborhood of g are sorted in increasing order and the *unnormalized enrichment score* ES_g is calculated according to Equation 4.2 (D). To normalize by local subnetwork size, random samples of equal size to A_g are drawn from all input p-values and a ES_{BG} value is computed for each of them (E). The distribution of ES_{BG} values is then used to estimate the enrichment p-value p^*_g, according to Equation 4.3 (F). FC denotes Fold Change (log_2) between two conditions.

4.2 Application to dengue data

We used LEAN for network analysis of dengue genotyping and transcriptomic data to search for associations with severe dengue.

4.2.1 Application to genotyping data

As input, we used gene-level p-values generated in Chapter 2, along with the STRING interaction network v9 (with the top-scoring 250,000 interactions, corresponding to an interaction confidence score of 0.637 or better) [Franceschini et al., 2013].

No significant network was found by LEAN at a significance level of $\alpha = 0.05$ on this genotyping dataset. This may have been because of the incomplete knowledge included in the network, or because of the imperfect functional mapping, as discussed in Chapter 2. The lack of a strong genetic signal in the cohort may likewise explain this result, potentially because a larger cohort would be needed to unravel complex relationships. Moreover, our input interaction network may lack important interactions, or include too many interactions that are irrelevant for dengue severity. The initial assumption of LEAN may also be inappropriate to dengue biology: subnetworks of the form of a gene and its direct neighbors may not aggregate the genetic signal in the right way. Furthermore, as we have no strong evidence that dengue severity is genetically determined, the variability explained by the genetics alone may not be large enough to be detected. Environmental factors, such as previously encountered pathogens, play a big role in dengue pathogenesis, as explained in Chapter 1. We were able to further examine this possibility using data that integrates the influence of these environmental factors, such as gene expression data.

4.2.2 Application to gene expression data

Data

I analysed expression in an *in vitro* experiment on monocytes from 11 patients from Thailand. For each of these patients, we have mRNA array-based gene expression measures of 70,524 transcripts performed using the HTA2 Affymetrix microarray. Expression is available under two experimental conditions: before infection by dengue virus, and after.

As explained in Chapter 1, after infection, in most people, dengue virus would multiply fast

in dendritic cells, causing high viral load. Ḃut some patients, are able to better resist to the infection, and their viral load stays low. The 11 patients comprised:

- 5 patients with high viral load after dengue virus infection, and
- 6 patients with low viral load after dengue virus infection.

LEAN analysis was performed to explore the molecular basis of the difference in reactions between these two subgroups.

Analysis

I have compared infected low-viral load versus infected, high-viral load samples, since, in non-infected samples, I observed no difference in between the two groups.

I first performed a Wilcoxon test, a non-parametric equivalent of the t-test, based on ranks, for all transcripts. This test does not require the assumption of normality, which, in turn, was impossible to test, given the small sample size. Moreover, it is more robust, i.e., less likely to indicate significance because of the presence of an outlier. No test turned out as statistically significant: among the 46 914 transcripts tested, none had a p-value that was lower than 0.5 after Benjamini-Hochberg multiple testing correction. (I also tested whether the result would change with a t-test. The same absence of significant results was observed.) I then used the input p-values and the same network as for genome data and performed LEAN analysis.

Results

Applying LEAN resulted in 352 local subnetworks being significant with a q-value of 0.05. The list of these genes appears in Appendix A.1. I then performed enrichment analysis of these genes using GSEA (described in Chapter 2). As background sets, I used the "hallmark" gene set from the MSigDB database, and C7, a set of immunological signatures of differentially expressed genes under different immune-specific perturbations. The complete

list of enriched sets can be found in Appendix B.1 for the background hallmark gene set, and the top 100 enriched immunological gene signatures from the background set C7 can be found in Appendix B.2. With both background sets we obtain results highly enriched in immunological responses. For the hallmark dataset as background, Table 4.3 presents an extract of most significantly enriched categories.

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	51		$4.18\ e^{-63}$	$2.09\ e^{-61}$
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	37		$1.71\ e^{-53}$	$4.29\ e^{-52}$
HALLMARK_OXIDATIVE_PHOSPHORYLATION [200]	Genes encoding proteins involved in oxidative phosphorylation.	36		$9.91\ e^{-39}$	$1.65\ e^{-37}$
HALLMARK_MYC_TARGETS_V1 [200]	A subgroup of genes regulated by MYC - version 1 (v1).	31		$2.32\ e^{-31}$	$2.89\ e^{-30}$
HALLMARK_DNA_REPAIR [150]	Genes involved in DNA repair.	27		$2.7\ e^{-29}$	$2.7\ e^{-28}$
HALLMARK_E2F_TARGETS [200]	Genes encoding cell cycle related targets of E2F transcription factors.	24		$8.35\ e^{-22}$	$6.96\ e^{-21}$
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]	Genes regulated by NF-kB in response to TNF [GeneID=7124].	21		$5.09\ e^{-18}$	$3.64\ e^{-17}$
HALLMARK_APOPTOSIS [161]	Genes mediating programmed cell death (apoptosis) by activation of caspases.	18		$4.07\ e^{-16}$	$2.55\ e^{-15}$
HALLMARK_G2M_CHECKPOINT [200]	Genes involved in the G2/M checkpoint, as in progression through the cell division cycle.	16		$3.35\ e^{-12}$	$1.86\ e^{-11}$
HALLMARK_ADIPOGENESIS [200]	Genes up-regulated during adipocyte differentiation (adipogenesis).	14		$4.49\ e^{-10}$	$2.04\ e^{-9}$
HALLMARK_INFLAMMATORY_RESPONSE [200]	Genes defining inflammatory response.	14		$4.49\ e^{-10}$	$2.04\ e^{-9}$

Figure 4.3: Top GSEA results using the hallmark dataset as background.

Interferon gamma response appears as most significantly enriched (False Discovery Rate (FDR) q-value of 2.10^{-62}). The second most enriched gene set is the interferon alpha response (FDR q-value of 4.10^{-52}). As explained in Chapter 1, interferons are involved in inducing inflammation, in the first reaction to infection. Other immunologic categories include TNF-alpha signaling via NF-kB. Non-directly related groups include genes implicated in genesis of adipose tissues. Also this result is consistent with prior knowledge: dengue severity is known to be associated with the quantity of lipoproteins (LDL and HDL) in blood [Biswas et al., 2015]. Other gene sets are related to apoptosis and more general cellular functions: MYC- and E2F-related groups, apoptosis, DNA repair, G2M checkpoint etc. These may be differentially expressed because of the lysis of infected cells. Table 4.4

presents an extract of the genes that fall into the most over-expressed categories.

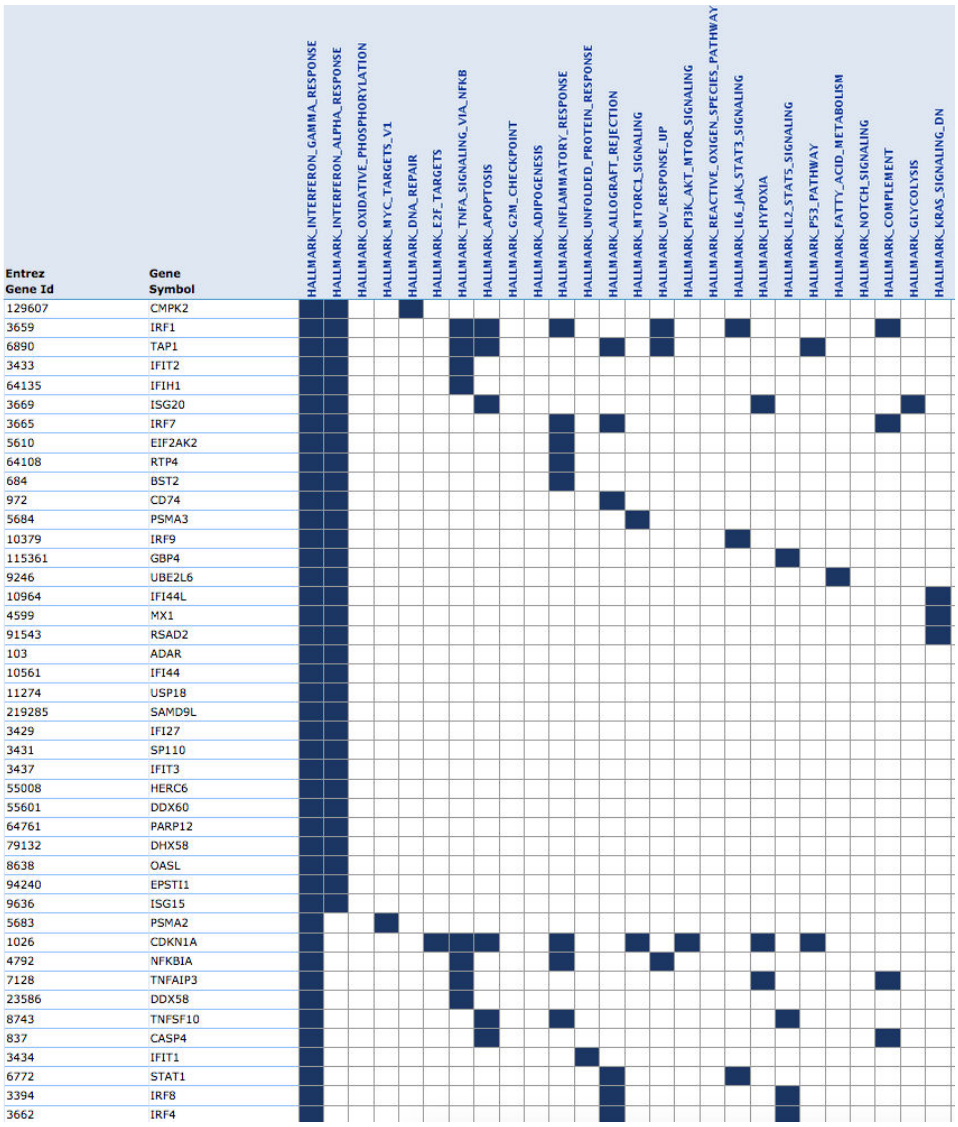
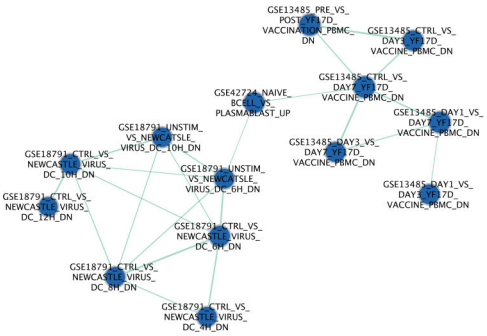


Figure 4.4: Extract of genes in top GSEA results using the Hallmark dataset as background.

Many of these genes have previously been associated with dengue severity in gene expression analyses, such as interferon inducible genes, OAS family gene OASL, TNF-kB family genes... [Coffey et al., 2009].

When using the more specific immunological signatures dataset as background, many gene sets have very significant enrichment p-values, similarly to the hallmark dataset. Since

there are over 100 enriched sets, I used the Enrichment Map Cytoscape App that allows connecting sets that share many genes [Merico et al., 2010]. The largest connected groups include genes that are upregulated in response to virus (Figure 4.5), and in response to lipopolysaccharides (LPS), long molecules on the surface of gram-negative bacteria known to induce a strong inflammatory reaction via interferons and NF-kB (Figure 4.6).

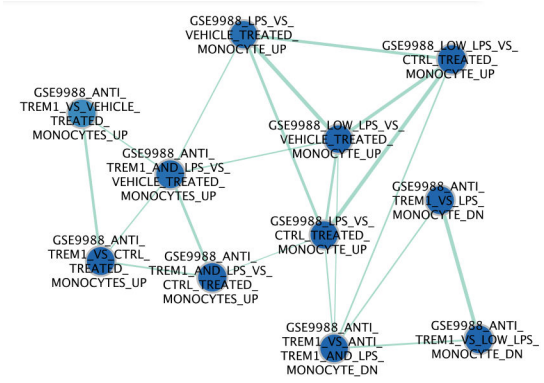


(a) Network of gene sets

	Gene
1	IFI44L
2	IFIT3
3	PARP9
4	DDX58
5	XAF1
6	IFIH1
7	DDX60
8	IFI44
9	RTP4
10	IFIT1
11	EPSTI1
12	IRF9
13	OASL
14	RSAD2
15	MX1
16	IFIT2

(b) Example of top genes in the gene set GSE13487

Figure 4.5: Gene sets upregulated during reaction to virus. (a) Gene sets that are signatures of genes upregulated during a specific viral infection. An interaction between two genes represents overlap between gene sets. (b) Most strongly differentially expressed genes that are present within the gene set GSE13487 that is part of the network in (a).



(a) Network of gene sets

	Gene
1	GBP1
2	GBP2
3	EGR1
4	IL18
5	TNFAIP3
6	REL
7	NFKBIA
8	CFLAR
9	PIGA
10	NFKBIZ
11	TNF
12	CSRNP1
13	DUSP1
14	TXN
15	MSH6
16	BTG2
17	PFKFB3
18	RIPK2

(b) Example of top genes in gene set GSE9988

Figure 4.6: Gene sets related to genes upregulated during inflammation. (a) Gene sets that are signatures of genes upregulated during a inflammation. Edges represents overlap between gene sets. (b) Most strongly differentially expressed genes that are present among most the gene set GSE9988 that is part of the network in (a).

4.3 Discussion

This chapter presents LEAN, an approach that I co-designed to aggregate omics data in the context of interaction networks. Here, I discussed the method itself, and its application to dengue genotyping and transcriptomic datasets. LEAN is able to compute best-scoring subnetworks and their empirical p-value, without relying on any user-tunable parameter, and without size bias, such as the one described in Chapter 3. It achieves this goal by only considering very specific subnetwork structures: a genes and its direct neighbors. The disadvantage of such a constraint may be that it is less powerful to identify statistical signals across gene sets that are connected, but not strongly interconnected between themselves. An extreme example of such a set is a linear pathway, where each node has only two connections to other members in the pathway, except from the extreme nodes that have

only one connection. Nevertheless, these gene structures that have a central node make a first step in aggregating signal and are much easier to interpret, since we can start by analyzing central nodes that may play a central role in the network. Moreover, LEAN needs to only explore one network per node, therefore decreasing greatly the space of networks to explore, compared to an algorithm such as jActiveModules (*cf.* Chapter 2).

Our application of LEAN to dengue disease generated diverse results: we found no significant results for our genomic data. For the transcriptomic data, we started by performing a test for differential expression for each individual transcript. The absence of low p-values during this test may well be due to the very small sample size, compared to the strength of the biological signal we can expect. By aggregating gene signals using LEAN, we were able to find sets of genes that were highly enriched in immune-related functions. Most of them are related to the current knowledge about the disease, reassuring us in that most results represent a true biological signal, rather than noise. This suggests that LEAN may indeed be capable of pinpointing specific genes in biologically relevant processes.

A next step of this analysis would be to generate new hypotheses for the differences in viral load, based the LEAN results, and to validate them experimentally. To generate these hypotheses, we need to focus on specific gene sets, groups of related gene sets, or on specific genes from gene sets, and interpret their role in the experiment. Once specific genes of interest are identified, it would be natural to consider their network neighborhood. Generating these hypotheses would therefore require close interactions with researchers specialising on dengue, or immunologists.

Another next question of interest is: Given the strength of the signal in this gene expression dataset, is it possible to create a biomarker that is able to predict dengue severity early on in the disease and direct hospital resources towards severe patients? The next chapter represents an attempt at answering this question.

Chapter 5

A machine learning approach to analyse dengue transcriptomic data

Contents

Here, I aimed to explore another approach to tackle complexity beyond single genes in biological data. I specifically aim to search for a multiple-gene biomarker that predicts the severity of the future reaction to dengue infection in patients, based on their blood transcriptomes at the earliest possible clinical stage, i.e., when they enter the hospital. Such a biomarker may ultimately be used to help doctors reliably distinguish between patients who can be sent home and those who are at risk to develop severe dengue, and need to be monitored in hospital. A second objective is to study genes included in the biomarker as starting points for deeper exploration and understanding of severe dengue. We will here first define the concept of a biomarker, then I will present the method that we developed for biomarker search, and finally, I will present the application of this method to gene expression data. At the time of this writing, this last part has been submitted as a journal article.

5.1 Biomarker: A definition

The term "biomarker" is a portmanteau of "biological marker". In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [Biomarkers Definitions Working Group, 2001].

Disease-related biomarkers either indicate of whether the patient is ill (diagnostic biomarker), the probable effect of treatment (predictive biomarkers), or how a disease may develop (prognostic biomarker) [Tezak et al., 2010].

In the context of the following analysis, we will also employ the term "gene signature", which is an other commonly used expression to designate a disease-related set of genes. Furthermore, we avoid the term "prognostic biomarker", as this might be considered an overstatement—some of the patients already had symptoms of severe dengue when entering the hospital. We will use the wording "biomarker that detects severe dengue" instead.

5.2 Classification through ensemble monotonic regression

Motivated, in part, by the work presented in the previous chapters, we aimed to develop a method that:

- would generate biomarkers with a small and controllable number of features,
- whose features we will be able to interpret biologically,
- that is able to generate linear and non-linear boundaries between phenotype classes,
- would generate a biomarker containing a stable feature set,
- allows fast enough algorithms to deal with a set of tens of thousands of transcripts,
- is suitable for datasets of tens of patients.

One classically used model for binary phenotypes is Lasso logistic regression. Nevertheless, Lasso can only generate linear boundaries between cases and controls. We were also interested in being able to find logical relationships such as: "if we have a high/low expression of transcript 1 AND/OR a high/low expression of transcript 2", the predicted phenotype is severe. Such relationships have been shown to exist in the biology of cancer [Iorio et al., 2016]. In modelling disease state as a function of two transcripts, an "AND" rule could capture, for instance, the role of a pair of key transcripts in two alternative pathways for a hypothetical physiological function lacking in severe patients. Severe patient status would then be correlated with low expression in both transcripts. In an "OR" rule, a low level of either transcript could correspond to a critical malfunctioning protein complex in severe disease. An interesting choice of a regression model that was able to find linear and non-linear interactions, including the logic functions above, and be fast enough to deal with all the features appeared to be monotonic regression. The only hypothesis made is monotonicity of the outcome: for a given transcript it can either be "the lower the expression the more severe the phenotype", or "the higher the expression the more severe the phenotype".

A mathematical definition of monotonicity is the following:

Definition [Isotonic, monotonic function]

A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $x = (x_1, \ldots, x_n) \mapsto f(x)$, is *isotonic in* x_i if f is an increasing function in x_i, *i.e.,*

$$\forall \Delta \geq 0, \forall x \in \mathbb{R}^n : f(x_1, \ldots, x_i + \Delta, \ldots, x_n) \geq f(x_1, \ldots, x_i, \ldots, x_n).$$

f is called *monotonic* in x_i if f is either increasing or decreasing in x_i, *i.e.*, f or $-f$ is isotonic in x_i. f is called *monotonic* if f is monotonic in all x_i. Figure 5.1 presents an illustration of a monotonic function in two variables (or two-dimensional monotonic regression).

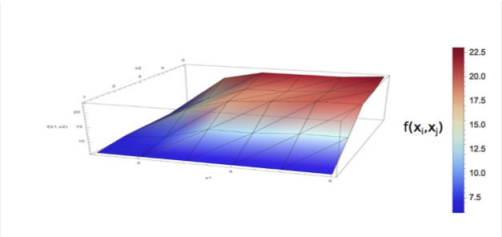


Figure 5.1: Illustration of a monotonic function of two variables.

Two-Dimensional monotonic regression for classification

To model the relation of a combination of transcript levels to the phenotype, we first use a monotonic function of two variables that best fits our training data, as measured by the L_1-norm. In our case, we have only two phenotypes: severe and non-severe dengue, which we encoded as 0 and 1, respectively. Thus, the optimal fit according to the L_1-norm is a monotonic function that minimizes the number of misclassified patients on the given data. As in the case of linear regression, from a fit by a regression function of the training dataset, we define a boundary between separating cases and controls by minimising an error function as explain in Chapter 1. This boundary enables us to classify previously uncharacterised patients.

The two-dimensional (2D) monotonic regression algorithm that I used fits a monotonic regression model of two transcripts to best predict the phenotypes of patients in the training

dataset, and estimates the performance of such a 2D model by leave-one-out cross-validation (LOOCV). (For the definition of leave-k-out cross-validation, see Chapter 1.) The algorithm was implemented mainly by Benno Schwikowski in C, and Mathematica. I used the C code for biomarker discovery on a cluster with 315 cores. The algorithm takes less than 30 minutes to run, being for the first time, fast enough to evaluate all possible pairs of transcripts using a recent algorithmic improvement [Stout, 2012]. The 2D version of the algorithm enables to take into account two transcripts to predict the phenotype. For a complex disease, it is an improvement compared to 1D monotonic regression, but is still limited. Therefore, from these results, I created the final biomarker on a standard personal computer using Python and Mathematica as follows.

Ensemble monotonic regression

To adapt to diseases that require many transcripts for prediction and to add robustness to the biomarker, I combined transcript pairs that had the smallest LOOCV error estimate (further referred to as *top pairs*) in a single ensemble biomarker: for a new patient p, the final phenotype is the proportion of top pairs that have predicted p as severe dengue. The exact number of pairs to include in the biomarker is determined statistically by an other round of LOOCV (*cf.* Figure 5.2 for the full pipeline of the method). One can also visualize pairs and is able to decide how many/what pairs to include in the ensemble biomarker. We therefore allow the user to choose a number of pairs that is smaller, than the one generating the optimal performance estimate by LOOCV, and therefore the user may trade off the complexity of the model and the estimated performance of the model. This option may be useful when trying to generate a biomarker with a small number of transcripts. The following analysis presents the successful application of this algorithm to detect severe dengue. The following part it a slightly adapted version of the article, that has been submitted to review in July 2017.

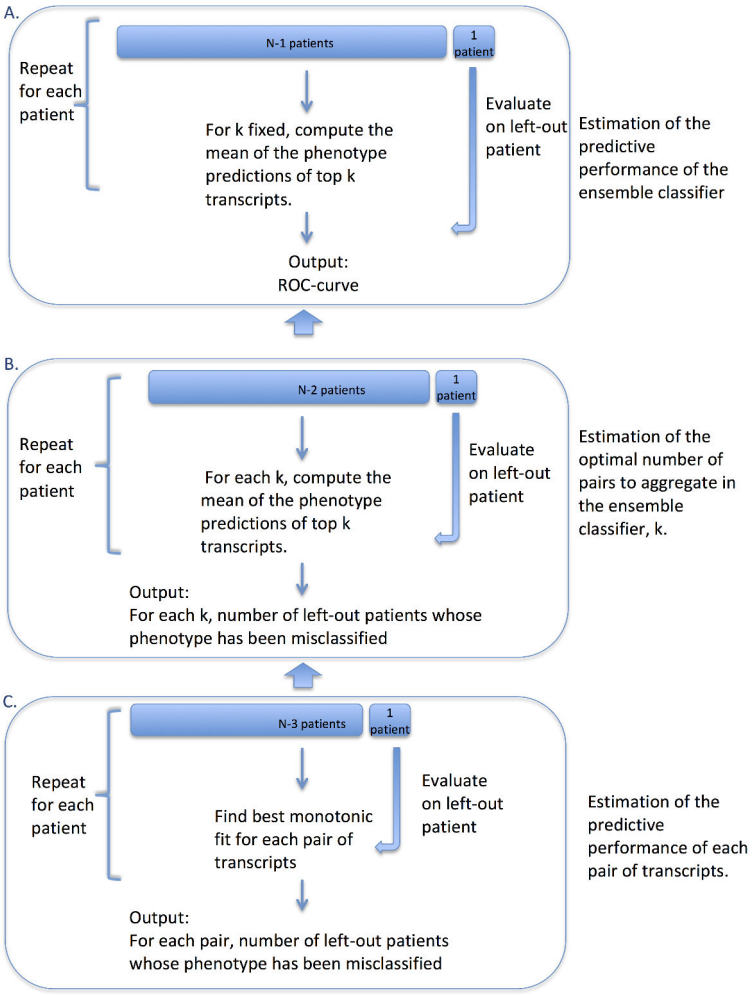


Figure 5.2: Whole pipeline of the estimation algorithm for N patients of the training cohort. It has three nested loops of LOOCV: We first do leave one patient out for the final evaluation (Fig.A), then we leave one patient out to estimate k_{opt}, the optimal number of pairs to include in our classifier (Fig.B), then we leave one patient out to estimate the predictive performance of each pair (Fig.C). Once these leave-outs are done, we start by first evaluating pair performance (Fig.C), then estimating k (Fig.B), and finally getting the final performance of our total classifier (Fig.A). Once the estimation of the performance finished, the final classifier is obtained by rerunning Fig.C on all the N patients and including k_{opt} best pairs on our classifier.

Important note : In the following section, we do not have separate validation data. We have a "training cohort" that we use to estimate parameters via leave-one-out cross-validation and three "test cohorts". However, the article from this chapter is written for a biomedical audience. We adopt their variation of the terminology, where "validation cohort"(instead of "test cohort") is used to denote independent data on which the classifier is evaluated without modifying any parameters.

5.3 A blood biomarker detecting severe disease in young dengue patients at hospital arrival

Authors: Iryna Nikolayeva, Pierre Bost, Isabelle Casademont, Veasna Duong, Fanny Koeth, Matthieu Prot, Urszula Czerwinska, Sowath Ly, Kevin Bleakley, Tineke Cantaert, Philippe Dussart, Philippe Buchy, Etienne Simon-Lorière, Anavaj Sakuntabhai, Benno Schwikowski

5.3.1 Summary

Background

Early detection of severe dengue can improve patient care and survival. To date, no reliable single-gene biomarker exists. We hypothesized that robust multi-gene markers exist.

Methods

We performed a prospective study on 438 Cambodian dengue-suspected patients, aged 4 to 22. We analyzed transcriptomic profiles of peripheral blood mononuclear cells (PBMCs) collected on the first day of hospital admission for 42 of these patients using microarrays. We developed a novel biomarker discovery approach that controls the number of genes included, and captures non-linear relationships between transcript concentration and disease severity. For evaluation, we estimated the predictive performance of the biomarker on previously

uncharacterized 22 PBMC samples from the same cohort using qRT-PCR and 32 whole-blood microarray transcriptomes from an independent cohort.

Findings

We identified an 18-gene biomarker for detecting severe disease in dengue patients upon hospital admission with a sensitivity of 0.93 (95% CI: 0.80-1.00) and a specificity of 0.67 (95% CI: 0.49-0.84) with a total area under ROC-curve (AUC) of 0.86 (95% CI: 0.75-0.97). The signature was validated on previously unseen data from 22 patients from the same cohort, with an AUC of 0.85 (95%CI: 0.69-1.00). In addition, it was validated on whole blood transcriptomic data from an independent cohort of 32 patients with an AUC of 0.83 (95%CI: 0.68-0.98).

Interpretation

Based on its robust performance, this biomarker could detect severe disease in dengue patients upon hospital admission, or even for prognosis if confirmed in further studies. Furthermore, its genes offer new insights into severe dengue mechanisms.

5.3.2 Introduction

Dengue is the most widespread mosquito-borne viral infection worldwide. Currently, 40% to 50% of the world population lives in areas at risk for dengue virus transmission.[WHO, 2017] If the majority of dengue cases are uncomplicated, it is estimated that each year 500,000 cases, mostly children, progress to severe dengue (SD) and require hospitalization. According to the World Health Organization (WHO), about 2.5% of those affected by severe dengue requiring hospitalization are still dying from complications.[WHO, 2017] The recent explosive spread of the related Zika virus might further increase this burden. Indeed, the complications associated with severe dengue are more common after secondary infection

than after primary infection,[Halstead, 2014] and recent studies both in vitro and in vivo have highlighted the potential of anti Zika immunity to trigger dengue enhancement.[Stettler et al., 2016] As recently highlighted by the WHO, robust and early detection of severe dengue, along with access to proper medical care, would not only decrease the fatality rate down to 1%, but also reduce health care costs and economic burden of the disease.[WHO, 2017]

While diagnosis methods for dengue infection are well established, there are no prognostic tests to help the clinician evaluate the risk of progressing to severe dengue. A number of biomarkers that use clinical variables for detecting severe cases of dengue infection have been proposed, both for adults and/or children.[Tuan et al., 2017, Lee et al., 2016, John et al., 2015, Soundravally et al., 2015, Thanachartwet et al., 2015, Pang et al., 2016] Nevertheless, none of the biomarkers we found in the literature have been replicated on independent datasets. In addition to these studies, others have aimed to identify molecular biomarkers, based on either mRNA expression, or on protein or cytokine levels. A number of genome-wide expression profiling studies have also been performed in Nicaragua, Cambodia, Thailand and Vietnam.[Kwissa et al., 2014, Devignot et al., 2010a, Popper et al., 2012, Hoang et al., 2010, Simmons et al., 2007a] Every study uncovered differentially expressed genes associated with severe dengue. Many of these genes have functions associated with innate immunity, vascular permeability, coagulation, neutrophil-derived antimicrobial resistance, inflammation, and lipid metabolism. However, their capacity to detect severe cases among dengue patients was not evaluated, [John et al., 2015, Soundravally et al., 2015, Thanachartwet et al., 2015, Kwissa et al., 2014, Devignot et al., 2010a, Popper et al., 2012, Hoang et al., 2010] or they exclude children.[Pang et al., 2016] Dengue is known to be a complex disease. To address this, a recent review suggested the study of combinations of molecules for the detection of severe cases.[John et al., 2015] To this end, Nhi et al. identified 19 plasma proteins exhibiting significantly different relative concentrations ($p - value \leq 0.05$) on 16 patients (6 severe dengue, 10 non-severe).[Nhi et al., 2016] Among them, a combination of antithrombin III and angiotensin had strong power to detect the 6 severe dengue patients (area under the ROC curve (AUC) = 0.87). Pang et al. developed

a biomarker combining transcript, protein and clinical markers, mostly linked to innate immunity and coagulation, that was able to detect patients with warning signs and needing to be hospitalized with sensitivity of 96% and specificity of 54.6% on a validation cohort.[Pang et al., 2016] However, these studies share a common drawback: none of the biomarkers have been replicated on an independent cohort. We hypothesized that a simple combination of a small number of gene expression markers may be robust enough to establish reproducible detection of severe cases among newly admitted dengue patients. With this in mind, we attempted to develop a biomarker discovery algorithm; one that allowed for not only linear but also more general monotonic relationships between features, meaning more complex, but still easily interpretable, relationships between genes.

Our underlying goal was to identify a biomarker able to detect severe cases from blood samples taken upon dengue patient admission to hospital. We conducted a prospective study in Cambodia of patients admitted to hospital with suspected dengue infection. Severe dengue cases were identified according to the WHO 2009 criteria using data at admission and during hospital stay. Our data consisted of gene expression profiles of peripheral blood mononuclear cells (PBMCs) on the date of admission. A PBMC is any peripheral blood cell having a round nucleus. These include important immune players such as lymphocytes (T cells, B cells, NK cells) and monocytes, but exclude red blood cell, platelets and granulocytes (neutrophils, basophils, and eosinophils). To control for the number of genes in the biomarker, and identify monotonic relationships between transcript concentrations and disease severity, we developed a new biomarker discovery approach. Using this, we identified an RNA biomarker of 18 genes in PMBCs that could detect severe dengue cases. We were able to replicate these results on previously unseen PBMC samples and whole blood samples taken using different technological platforms. From the known functions of these genes, we obtained new insights into the pathophysiology of severe dengue.

5.3.3 Research in context

Evidence before this study

We searched the PubMed database for "dengue"[Title] AND (severe OR severity OR shock) AND (risk OR biomarker) AND (human OR patients) without any date restrictions. Even though most severe dengue cases occur in children, none of the biomarkers in the resulting literature that included samples from children had stated sensitivity and specificity on an independent cohort.

Added value of this study

Young patients are particularly at risk for severe dengue infection in endemic regions. Our study presents the first independently validated molecular biomarker detecting severe dengue in this patient group with stated measures of specificity. Estimates of predictive performance on two independent cohorts were stable across biological and technical variation, and had an AUC (area under ROC curve) ranging from 0.83 to 0.85.

Implication of all the evidence

This study provides the first evidence that a well-performing molecular biomarker for detecting the severe form of the disease in young dengue patients across different technical conditions and blood cell subtypes is possible. The novel non-linear model underlying the biomarker is flexible enough to discover complex gene-gene interactions, yet simple enough to be represented visually. Our analysis of the included biomarker genes confirms several previous findings, as well as suggests new biological processes that may help understand severe dengue.

5.3.4 Materials and methods

Population studied

We conducted a prospective study in the Kampong Cham referral hospital Cambodia during a 3-year period (2011-2013). Patients suspected of dengue infection were invited to participate in the study. Dengue infection was confirmed by positive RT-PCR and/or positive dengue NS1 antigen detection. Three blood samples were collected: (i) shortly after hospital admission during the febrile acute phase, (ii) at the time of defervescence, and (iii) during the convalescent phase at the time of hospital discharge. In this study, we used only the transcriptome of blood samples collected shortly after hospital admission for both the microarray training set patients, and qRT-PCR validation set patients. This corresponded on average to the third day after onset of fever (Figure 5.3). We focused our analysis on samples of secondary DENV-1-infected patients that were judged to be of sufficient quality and quantity for this analysis, which resulted in 42 samples for microarray analysis and 22 samples for qRT-PCR analysis. Blood samples were processed as follows: plasma was used for dengue confirmatory diagnostic including serology and molecular diagnostics, as described elsewhere[Duong et al., 2015], while blood clot and PBMC were kept for later analyses. For this PBMC cohort, diseased severity was classified according to the 2009 WHO criteria using clinical and biological data recorded at admission and throughout the entire hospitalization period.[WHO (World Health Organisation), 2009] For the independent whole blood microarray cohort, disease severity was classified according to the description in the Section "Biomarker discovery" below.

Ethics statement

The study was approved by the Cambodian National Ethics Committee for Health Research (approval no. 087NECHR /2011 and no. 063NECHR/2012). Before a participant's enrollment, written consent signed by the participant or by a legal representative for participants under 16 years of age was obtained.

	Clinical cohort			Devignot et al. cohort
	Entire clinical cohort	PBMC microarray subcohort Training set	PBMC qRT-PCR subcohort Validation set	Whole blood microarray subcohort Validation set
Suspected dengue cases	438	42	22	32
Age	8·8±3·3	9·0±3·8	8·8±3·1	7·8±2·5
Sex: male-to-female ratio	0·49	0·48	0·40	0·34
Mean day of blood sampling after onset of fever	3·3±1·5	3·1±1·9	3·6±1·3	5·1±1·0
Confirmed dengue	316	42	22	32
DENV-1	265	42	10	4
DENV-2	15	0	6	3
DENV-3	0	0	0	16
DENV-4	28	0	6	1
Unknown	8	0	0	8
Secondary infection	183	42	22	32
Classification according to WHO 2009 criteria				
Non-severe dengue, with or without warning signs	236	27	15	14
Severe dengue	80	15	7	18

Figure 5.3: Patient characteristics. We present all the cohorts used: our clinical cohort with training and validation subcohorts, as well as the validation cohort from [Devignot et al., 2010a].

RNA preparation, microarray hybridization and qPCR validation

RNA was extracted from PBMC stored in RNA protect cell reagent (Qiagen, Hilden, Germany) with a miRNeasy kit (Qiagen) and RNA quality checked on a BioAnalyzer 2100 (Agilent, Santa Clara, California). For microarray analysis of the training cohort, gene expression in PBMC was analyzed using Affymetrix Human Transcriptome Array 2 (HTA2) GeneChips. HTA2 chips were prepared, hybridized, and scanned according to the manufacturer's instructions. For qRT-PCR of the PBMC validation cohort, 200 ng RNA were reverse-transcribed with SuperScript VILO cDNA synthesis kit (Invitrogen, Life Technologies, Carlsbad, CA, USA), using a combination of random hexamer and Oligo(dT)12-18 primers. TaqMan Gene Expression Assays (Life Technologies) were used for each candidate gene according to the manufacturer's instructions. Relative expression was calculated with the $2^{-\Delta\Delta C_t}$ method, using beta glucuronidase (GUSB) as endogenous control for normalization and a calibrator sample as a comparator for every sample.

Plan for biomarker discovery

Our 18-gene biomarker was identified through an automated machine learning algorithm applied to microarray transcriptomes of the PBMC training cohort, leading to an initial assessment of its performance via rigorous cross-validation. After applying necessary quantile normalization (Section 5.3.9), we evaluated this biomarker on two previously unseen datasets (Figure 5.4).

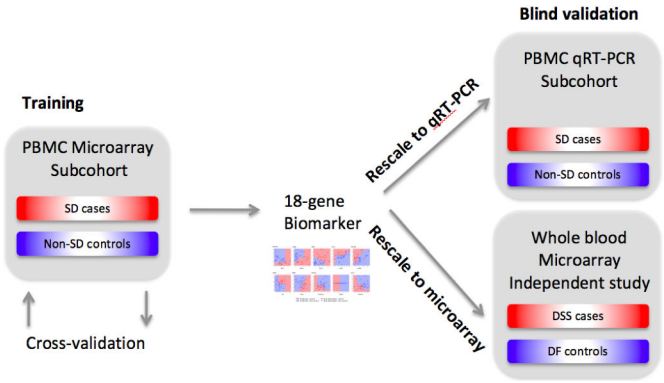


Figure 5.4: Flow diagram for the discovery and validation of the severe dengue (SD) diagnostic biomarker.

The first validation dataset consisted of 22 unseen patients (7 severe dengue, 15 non-severe) from the PBMC validation cohort, whose gene expression was measured using qRT-PCR. As the IGKC transcript found in the 18-gene biomarker was expressed at undetectable levels in the PBMC validation cohort, its levels were substituted with the measured levels of its partner PPBP in the PPBP-IGKC gene pair. The second validation dataset was an independent, publicly available, Cambodian whole blood dataset, selected for its large size and high quality.[Devignot et al., 2010a] It consisted of whole blood transcriptome data from 48 dengue-infected patients. At the time of that study, phenotype was still established according to the 1997 WHO classification: DSS (Dengue Shock Syndrome), DHF (Dengue Hemorrhagic Fever), and DF (Dengue Fever).[WHO (World Health Organisation), 1997] To make phenotype data comparable, we reclassified the disease severity as well as possible in terms of the 2009 WHO classification. We considered all 18 DSS patients as severe dengue, and all 14 DF patients as non-severe, considering that DF patients that are reclassified

as severe dengue in the 2009 WHO classification are rare. DHF patients could not be classified without additional clinical information that was unavailable to us, and were thus excluded.

Machine learning methodology

Our biomarker was created using a machine learning approach based on monotonic regression on a training cohort as explained in section 5.2. Briefly, new predictions made by the biomarker are based on 0/1 (non-severe/severe) predictions (votes) derived from pairs of transcripts in the biomarker. Measured concentrations for any given transcript pair are turned into a binary vote using a two-dimensional monotonic function,[Stout, 2012] a generalization of a linear function that monotonically increases or decreases with the concentration of each transcript. The final prediction is "severe" if the mean of all votes is above the threshold t, and "non-severe" otherwise. The performance of individual transcript pairs on future patients is estimated using cross-validation. The resulting biomarker consists of a set of transcript pairs with unique transcripts having an optimal performance estimate. Using a permutation test, we then eliminated those genes that did not confer a statistical performance advantage over the performance of their partner alone. The resulting model represents a unique combination of lower- and higher-complexity features tailored towards the discovery of complex disease biomarkers. The monotonic model generalizes linear models. Nevertheless, the resulting features can still be visually and intuitively understood. Controlling the number of transcripts in the biomarker allows different trade-offs between performance, robustness, and assay cost (Section 5.3.9) To rescale the biomarker to the different measurement units of our validation sets, we mapped transcripts to genes and quantile-normalized the expression values (Section 5.3.9).

Performance evaluation

We summarized biomarker performance using the ROC curve, which consists of the different combinations of true and false positive rates that are obtained by varying the above

threshold t between 0 and 1. For the comparison with state-of-the art machine learning methods, we used the implementations from the Python sklearn [Pedregosa et al., 2012] package.

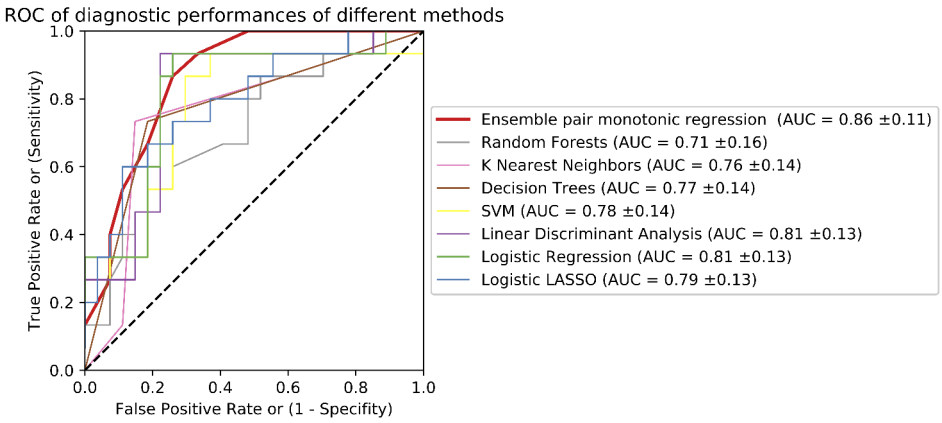
5.3.5 Results

We have identified an 18-gene biomarker that allows the detection of severe dengue from a blood sample taken from dengue patients upon hospital arrival. We evaluated the performance of the biomarker using two validation datasets. The first validation set was generated from PBMC transcripts of additional patients from the above cohort, which were quantified by qRT-PCR. The second validation set consisted of data from a whole blood transcriptome array from an independent, previously published study. [Devignot et al., 2010a]. The performance of our biomarker was estimated by cross-validation. We obtained AUC values of 0.86, 0.83 and 0.85 for the training set and the two validation sets, respectively. Twelve of the eighteen genes in the biomarker are immune-related (Table 5.1). Certain genes have already been associated with severe dengue.

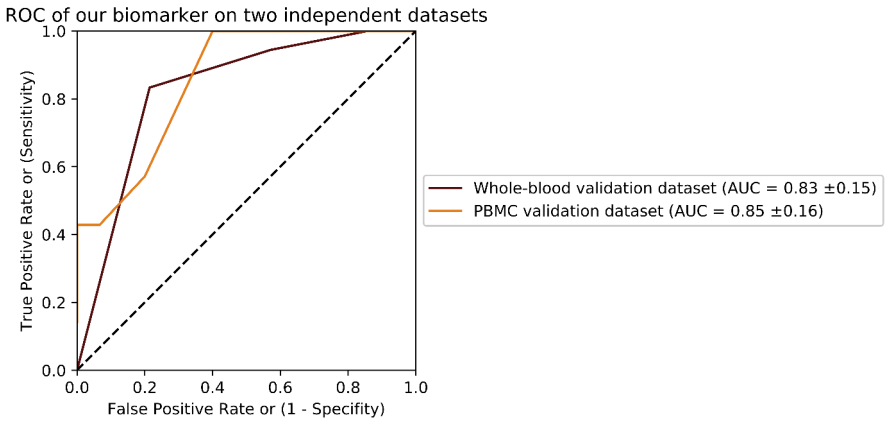
To determine whether the inclusion of a larger number of genes or the restriction to a linear state-of-the-art variable selection model would have increased classification accuracy, we estimated the performance of several well-known classification methods (Figure 5.5.a). Methods are presented in (Chapter 1, Machine learning). Though differences in performance did not reach statistical significance, our method gave the highest AUC. Moreover, logistic regression with a lasso penalty (logistic lasso), a state-of-the-art linear variable selection method, generated a classifier whose performance was not better than random on the PBMC qRT-PCR dataset (Section 5.3.9).

Gene	Gene Name	Description	Over/ Under (+/−) expressed in SD	Known link with SD in literature
E2F7	E2F transcription factor 7	Participates in various processes such as angiogenesis, polyploidization of specialized cells and DNA damage response. Acts as a negative regulator of keratinocyte differentiation.	+	
ENKUR	Enduring, TRPC channel interacting protein	Ca-mediated signaling	-	
ARG1	Arginase 1	Controls arginine metabolism in neutrophils, hence controlling NO production (iNOS pathway) moderator of T cell function.	+	[Hoang et al., 2010]
JUNB	JunB proto-oncogene, AP-1 TF subunit	Transcription factor involved in regulating gene activity following the primary growth factor response. Expressed in neutrophils. Part of the iNOS pathway.	-	
E2F7	E2F transcription factor 7	Participates in various processes, such as angiogenesis, polyploidization of specialized cells, and DNA damage response. Acts as a negative regulator of keratinocyte differentiation.	+	
MPO	Myeloperoxydase	Produced mainly by neutrophils. This enzyme produces hypohalous acids central to the microbicidal activity of neutrophils.	+	[Devignot et al., 2010b, Hoang et al., 2010]
LRP1	Prolow-density lipoprotein receptor-related protein 1	Endocytic receptor involved in endocytosis and in phagocytosis of apoptotic cells. Involved in the plasma clearance of chylomicron remnants and activated LRPAP1 (alpha 2-macroglobulin).	-	
PGD	Phosphogluconate dehydrogenase	Enzyme involved in the pentose phosphate pathway, hence producing more NADPH. NADPH is a cofactor used in anabolic reactions, such as lipid and nucleic acid synthesis, which require NADPH as a reducing agent.	+	
EGR3	Early growth response 3	This gene encodes a transcriptional regulator that belongs to the EGR family of C2H2-type zinc-finger proteins. It is an immediate-early growth response gene which is induced by mitogenic stimulation. The protein encoded by this gene participates in the transcriptional regulation of genes in controlling biological rhythm. It may also play a role in a wide variety of processes including endothelial cell growth.	-	
MGAM	Maltase-glucoamylase	This gene encodes maltase-glucoamylase that plays a role in the final steps of digestion of starch.	+	
HP	Haptoglobin	Binds free plasma haemoglobin, antimicrobial activity.	+	[Simmons et al., 2007b, Devignot et al., 2010b, Hoang et al., 2010]
MYB	Myeloblastosis proto-oncogene, transcription factor	Transcriptional activator, implicated in B cell lymphoma	+	
IGKC	Immunoglobulin kappa constant		+	
PPBP	Pro-platelet basic protein	Platelet-derived growth factor of the CXC family. It is a potent chemoattractant and activator of neutrophils and has anti-microbial properties.	-	
CD40L	CD40 ligand	This gene is expressed on the surface of T cells. It regulates B cell function by engaging CD40 on the B cell surface. A defect in this gene results in an inability to undergo immunoglobulin class switch and is associated with hyper-IgM syndrome.	-	
OX40L	OX40 ligand	Mediates adhesion of activated T cells to endothelial cells, expressed on antigen-presenting cells such as dendritic cells, endothelium, mast cells and NK cells.	-	
SDPR	Serum deprivation response	Participates to the formation of caveolae.	-	[Long et al., 2009]
TCF7	transcription factor 7 (T-cell specific, HMG-box)	This gene is expressed predominantly in T-cells and plays a critical role in natural killer cell and innate lymphoid cell development. The encoded protein forms a complex with beta-catenin and activates transcription through a Wnt/beta-catenin signaling pathway.	-	
ASAP2	ArfGAP with SH3 domain, ankyrin repeat and PH domain 2	The protein localizes in the Golgi apparatus and at the plasma membrane. The protein forms a stable complex with PYK2 in vivo.	-	

Table 5.1: Constitutive gene pairs of our biomarker. Genes are grouped into pairs (or singletons if the partner did not add any statistical advantage).



(a) Performance of different methods on the training dataset



(b) Performance of our biomarker on the validation datasets

Figure 5.5: Performance evaluation

a. Training: Performance of our biomarker compared to other methods. Performance estimates of state-of-the-art classification methods established by leave-one-out cross-validation on the PBMC microarray training set. Area under ROC curve (AUC) for each method is indicated with its 95% confidence interval.

b. Validation: ROC curves on independent datasets. To assess the performance of our biomarker, we blindly predicted the phenotype of new patients from the same cohort as our training set, and from an independently published cohort of whole blood samples.

Figure 5.6 provides a visualization of the models associated with the transcript pairs of the biomarker. Different monotonic functions capture different types of gene-gene interactions. For example, for the second pair of transcripts (JUNB and ARG1), patients have a severe phenotype when JUNB expression is high or ARG1 expression is low. For OX40L and

CD40LG, OX40L and CD40LG both are under-expressed in the severe patients. For EGR3 and MGAM, the lower the EGR3 expression, and the higher the MGAM expression, the more likely the patient is to be predicted severe.

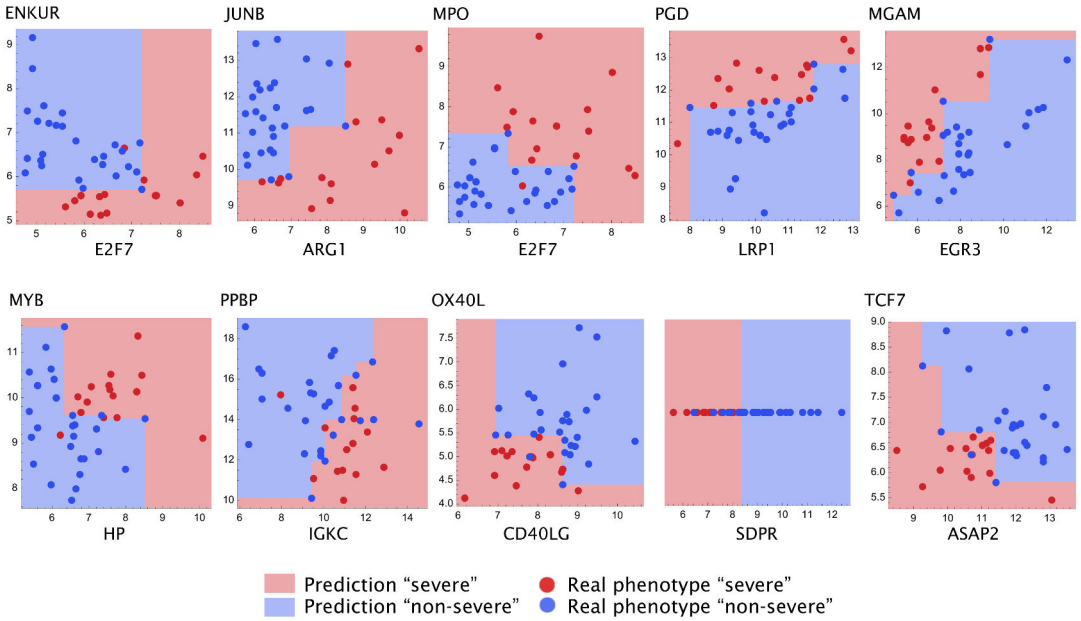


Figure 5.6: Visual representation of the biomarker. The biomarker is applied to a new set of transcript measurements by first making one prediction from each of the ten panels for each patient. Each such prediction is generated by reading off the panel's background color at the coordinates defined by the new transcript measurements. The final biomarker prediction is then made by comparing the resulting frequency of severe predictions against a threshold. For illustration, the panels show the points corresponding to transcripts from the PBMC training cohort. The biomarker can be applied to data on different measurement scales after quantile normalization.

5.3.6 Discussion

We have identified and independently validated a biomarker for the detection of severe cases among dengue patients from blood samples taken upon arrival at the hospital. Severity was defined according to the 2009 WHO dengue classification.[WHO (World Health Organisation), 2009] This 18-gene expression biomarker was built using PBMC samples of newly hospitalized Cambodian dengue patients using transcriptome microarrays. Our novel ap-

proach to biomarker discovery models linear and non-linear monotonic interactions between transcript levels with controlled complexity, and preserves interpretability and applicability to datasets of limited size. We performed a first validation of our biomarker by quantifying, using qRT-PCR, transcripts of previously uncharacterized PBMC samples from the same dengue season/cohort. The performance results remained stable compared to our original performance assessment. For further validation, we used a whole blood cohort from an independent public dataset.[Devignot et al., 2010a] To our knowledge, our results represent the first molecular biomarker for detecting severe cases in dengue patients with demonstrated high performance on independent datasets. The genes OX40L and CD40L that comprise our first gene pair are both under-expressed in severe cases (Figure 5.6). OX40L and CD40L are membrane proteins expressed by dendritic cells and by activated T cells, respectively, that are essential to mount an efficient adaptive immune response. OX40L binds to its co-receptor OX40 and allows T cells to survive after clonal expansion. Stimulation of B cells by T cells through CD40L is necessary for class switching and somatic hypermutation, and hence both genes are required to produce potent neutralizing antibodies (Figure 5.7).[Elgueta et al., 2009] In the context of dengue infection, OX40L has been shown to be down-regulated in human monocyte-derived dendritic cells after in vitro infection, supporting a role of the co-stimulatory molecule in dengue infection.[Gandini et al., 2011] In addition, we have observed a differential regulation of the expression of the OX40 signaling pathway in asymptomatic dengue cases compared to clinical cases (Duong, Simon-Loriere et al, in press). The role of CD40L in dengue infection is less clear; on one hand CD40L has been described as an enhancer of viral particle production by infected dendritic cells by providing survival signals,[Sun et al., 2016] but on the other hand CD40L is up-regulated in dengue specific CD4+ T cells and important for protection against the virus through an antibody-independent pathway.[Yauch et al., 2010] The second gene pair of our biomarker, ARG1 and JUNB, controls inflammation. Both genes are expressed in neutrophils and are known to regulate the production of reactive nitrogen species. ARG1 degrades the substrate of inducible nitric oxide synthase (iNOS).[Munder et al., 2005] JUNB transcriptionally regulates the expression of iNOS.[Ratajczak-Wrona et al., 2012] Hence, these genes together control the inflammatory status of the main blood component. Moreover, it has been found

that JUNB is a key transcriptional modulator of macrophage expression. It activates the expression of ARG1 in the presence of IL-4.[Fontana et al., 2015] The role of ARG1 in flavivirus infection has been extensively described; in the case of dengue, the production of RNS is required to inhibit viral replication during the early phases of infection. However an overproduction of RNS in the late phases of the disease leads to the inhibition of coagulation, leading to dengue-typical bleeding. ARG1 is therefore required to reduce the amount of RNS and bleeding during dengue infection.[Burrack and Morrison, 2014] This biomarker could be easily implemented in a clinical setting, and used sequentially or in combination to a dengue diagnostic test. Such a tool would allow more efficient patient triage, and close monitoring of individuals with high risk for severe disease, and would be especially useful in non-endemic regions where physicians might not possess extensive experience in dengue diagnosis and management. Indeed, this biomarker requires only a blood sample from the patient, and any technology that could measure the expression level of these 18 specific genes. Moreover, a recent large-scale study suggests that the concentrations of most proteins are linearly related to RNA concentration (with gene-specific levels).[Edfors et al., 2016] Thus, a protein-level implementation of our biomarker may potentially further ease its use, or allow its deployment in point of care settings.

In conclusion, we have presented a highly performing 18-gene biomarker that detects severe cases among dengue patients fast and objectively upon arrival at the hospital. Its performance was extremely stable on PBMC and whole blood samples, and across different technological platforms. A deeper understanding of the underlying biology, and how important parameters such as blood cell type, serotype, day of fever, and measurement platform impact the expected performance, will require dedicated follow-up studies. The potential of the marker as a prognostic marker for the early detection of risk of evolution towards severe dengue remains to be determined in further studies.

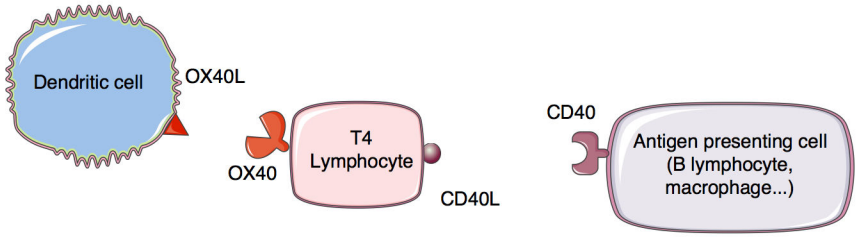


Figure 5.7: Activation of antigen presenting cells via OX40L and CD40L.

5.3.7 Author contributions

VD, PBu, AS designed the studies. VD, SL, PBu, PD collected the samples and clinical data. ESL, IC, MP and FK did the transcriptome quantification experiments. IN, BS and KB designed the methods. IN, UC and BS analyzed the data. PBo and TC interpreted the data. IN, PB, IC, VD, PD, ESL, AS and BS wrote the manuscript. All authors read the manuscript and approved its submission.

5.3.8 Role of the funding source

BioMérieux project and DENFREE consortium funding was used for data generation. IN was supported by Labex IBEID , the doctoral school Frontières Du Vivant and OpenHealth Institute.

5.3.9 Appendices to this chapter

Probe filtering

It has been found that appropriate independent filtering increases detection power for high-throughput experiments [Bourgon et al., 2010]. Thus, when possible, we used such filtering. Only transcripts with variance greater than 0.5 have been kept for the analysis. Moreover, since interpretability of our results was key, we have kept transcripts that had an Entrez

gene ID. This resulted in 2,653 transcripts being analyzed.

As to patient filtering, criteria are detailed on Figure 5.8

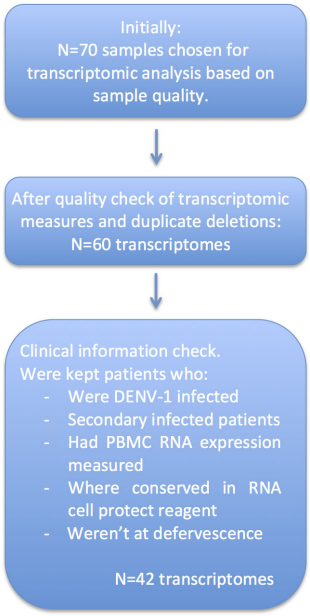


Figure 5.8: Description of transcriptome patient filtering.

Rescaling of new datasets

To be able to use the biomarker on new datasets, we need to make the transcript measures comparable in-between datasets. We quantile-normalised [Amaratunga and Cabrera, 2001] each validation dataset with our PBMC training dataset. For the PBMC validation dataset, since the measures came from relative qRT-PCR quantification, gene expressions where incomparable for different genes. Thus, the quantile-normalisation was done for each gene separately. More precisely: we first ensured ourselves that we have the same proportion of cases and controls in our training set that in the validation set. If the proportion of cases was lower(resp. higher), we duplicated a random cases(resp. controls) to equalise these proportions. Then, for:

- PBMC validation data: for each gene A, we ordered gene expressions of patients in

PBMC training dataset and in the PBMC validation dataset. This generated two ordered lists L_{train} and L_{other}.

- Whole blood validation data: Genes had already comparable measures in between themselves, due to the properties of transcriptomic arrays. We thus quantile-normalized the whole array taken together: We assumed that in reality the distributions of the gene expressions should be similar globally. Thus we pooled patients and genes together and ordered expression values in training dataset and application dataset. This generated 2 ordered lists L_{train} and L_{other}.

  Then, to the i-th value of the validation dataset we attribute the value in the training set with the index $i_{new} = Round(i * Length(L_{train})/Length(L_{other}))$

- PBMC validation dataset: the values in between genes were not comparable, thus we did the above normalisation for each gene separately instead of doing it once for the whole dataset.

Varying the number of pairs included in the biomarker

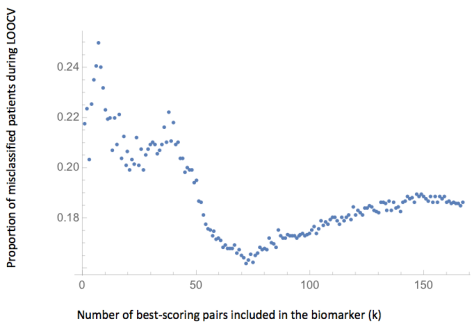


Figure 5.9: Estimating optimal k, the number of feature to include into the ensemble classifier, by calculating the proportions of mispredictions for each k.

We wanted to assess the impact of a simplification of the biomarker. Even though our optimal performance was obtained when using 74 different pairs of transcripts, the AUC

decreases by only 2.5%. Given the 95% confidence intervals of the AUC scores, this decrease is not significant.

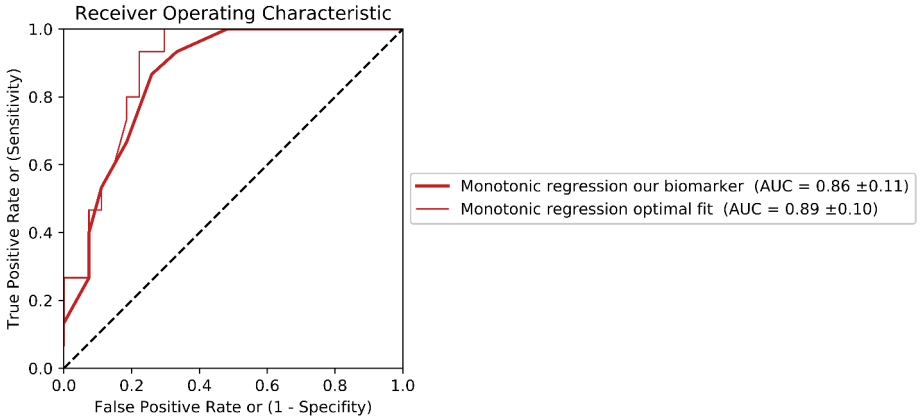


Figure 5.10: Impact of the simplification of the biomarker on the performance.

Performance of the logistic Lasso biomarker on the validation datasets

The optimal biomarker (for parameter lambda=1 s.e.), consisted of two genes, ARG1 and MPO. The AUC suggests that the resulting biomarker is not robust enough to reproduce its good performance on our validation PBMC dataset, even though it had a good performance on the independent whole blood dataset.

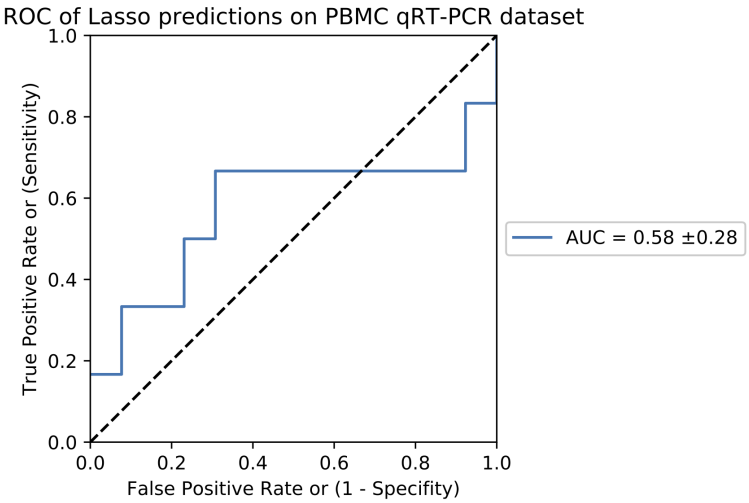


Figure 5.11: Performance of the logistic Lasso biomarker on the qRT-PCR dataset.

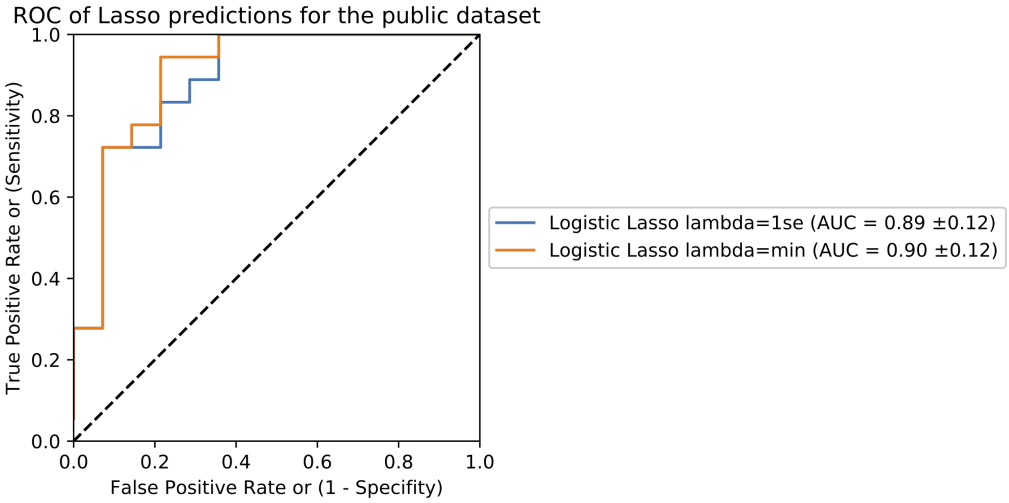


Figure 5.12: Performance of the logistic Lasso biomarker on the public dataset.

Additional information on genes in the biomarker

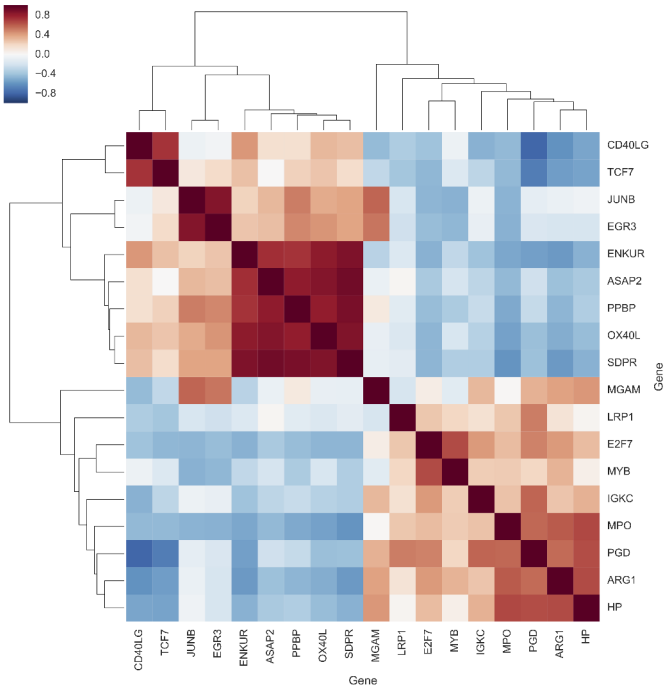


Figure 5.13: Correlation between genes included in the biomarker in the PBMC training dataset.

Gene	Transcript Probe ID
E2F7	TC12001756.hg.1
ENKUR	TC10001111.hg.1
ARG1	TC06000983.hg.1
JUNB	TC19001995.hg.1
E2F7	TC12002970.hg.1
MPO	TC17001727.hg.1
LRP1	TC12002396.hg.1
PGD	TC01000129.hg.1
EGR3	TC08002253.hg.1
MGAM	TC07000899.hg.1
HP	TC16002057.hg.1
MYB	TC06003069.hg.1
IGKC	TC02003395.hg.1
PPBP	TC04001282.hg.1
CD40L	TC0X000666.hg.1
OX40L	TC01003525.hg.1
SDPR	TC02002627.hg.1
TCF7	TC05002628.hg.1
ASAP2	TC02000046.hg.1

Table 5.2: Identifiers of transcripts in the biomarker from Affymetrix HTA2 array and associated gene names.

Chapter 6

Conclusion

Contents

6.1 Summary

Throughout this thesis, I have studied susceptibility to severe dengue through genotyping and transcriptomic data. Since dengue is a complex disease, I used approaches that allow to aggregate signal across many genes, based on pathways, interaction networks and machine-learning algorithms. I started by exploring genomic data from a recently published GWAS, which allowed to uncover associations between the two gene MICB and PLCE1, and severe dengue. By mapping SNP p-values to genes, we find additional significant p-values for several MHC genes (HLA-B, MICA and HCP5). The MHC contributes to processing and presenting antigens on the surface of infected cells in order to trigger the immune response. Network analysis of these genes thus leads to new results: The resulting network from the HumanNet interaction network is enriched in genes related to antigen processing and presentation via MHC class I. Additionally, it has a group of genes associated to the complement activation classical pathway, an alternative immune reaction pathway that our body uses to fight viruses. Moreover, the resulting network from STRING gene functional network is enriched in kidney development related functions (FOXC2, PLCE1, ASS1, POU3F3, PYGO1, and AGTR1 genes) among which are blood volume control and stimulation, and contraction of muscular tissue of capillaries and arteries via AGTR1. These functions are critical to avoid developing strong plasma leakage out of blood vessels and to avoid clinical shock. Plasma leakage and shock are included in the characterisation severe dengue. Therefore this data suggests, that there is a genetic predisposition to develop severe dengue depending on the alleles of genes related to blood volume control and stimulation, and contraction of muscular tissue of capillaries and arteries. These results need to be analysed with caution, since the result of jActiveModules does not give any measure of statistical significance. Nevertheless enrichment analysis and biological interpretation are coherent with what we know about dengue mechanisms and suggest that we observe is a true signal.

The variability of results and scores obtained by using input networks of different sizes brought us to study the jActiveModules scoring. In Chapter 3, we show that the score of

jActiveModules, as described in [Ideker et al., 2002], under the null hypothesis of uniformly distributed p-values, is biased when comparing scores of networks of different sizes for a fixed input network. By normalisation against size-specific null distributions of the generalised extreme value family, we derive a theoretical new, unbiased, score. This score function is computationally hard, and we outline our view of existing best practical options to avoid size bias.

In Chapter 4, we describe the design and use of a tool that does not suffer from the bias described in Chapter 3, LEAN. We use LEAN network to analyse GWAS data of Chapter 2, without finding any significant result. Motivated by the fact that environmental factors have a great impact on dengue outcome, we then analyse gene expression data and compare gene expression during an *in vitro* experiment that consists of injecting dengue virus into blood, determining the viral load of cells, and comparing samples that develop a high viral load with those that develop a low viral load. This analysis produces results highly enriched in different processes mostly related to inflammatory signalling, with a highest score for interferon alpha and gamma signalling.

In Chapter 5, we hypothesize that a combination of expressed mRNAs can help detect disease severity upon arrival at the hospital. We developed a machine learning method that is able to go beyond linear interactions, without using overly complex models to not overfit our learning dataset of 42 individuals. We evaluate the predictive performance of all monotonic relationships between all the pairs of transcripts and the phenotype, then we assemble best-scoring pairs to create an 18-gene biomarker. This biomarker predicts severe dengue with a sensitivity of 0.93 (95% CI: 0.80-1.00) and a specificity of 0.67 (95% CI: 0.49-0.84) with a total area under ROC-curve (AUC) of 0.86 (95% CI: 0.75-0.97). The signature was validated on previously unseen data from 22 patients from the same cohort using an other mRNA quantification technique, qRT-PCR, with an AUC of 0.85 (95%CI: 0.69-1.00). In addition, it was validated on whole blood transcriptomic array data from an independent cohort of 32 patients with an AUC of 0.83 (95%CI: 0.68-0.98). Our signature has the advantage of being easy to visualize, facilitating its interpretation. Interestingly, E2F and MYB are genes in common with the previous experiment in Chapter 4. As in

Chapter 2, we find that the antigen presenting process is important through the OX40L and CD40L pair. Both genes are membrane proteins expressed by dendritic cells and by activated T cells that are essential to mount an efficient adaptive immune response. Many other genes are linked to immuno-modulation via T/B-cell activation (MYB, IGKC, CD40L, OX40L, TCF7, ARG1), and neutrophils (ARG1, JUNB, MPO, PPBP).

6.2 Discussion

6.2.1 Biological results

Interestingly, genes identified by our genomic and transcriptomic analyses point to the same processes in severe dengue: immune processes implicating adaptive and innate immunity, especially related to the antigen presentation on the immune cell surfaces, appear to play a role in the pathogenesis of severe dengue. Our first network analysis of GWAS also shows an association with genetic predispositions to regulate vessel permeability and blood volume; nevertheless, the statistical significance of this association was not confirmed when using LEAN method. Chapter 5 moves one step beyond this by introducing a biomarker to improve early detection of severe dengue, especially by non-experienced doctors. Such a biomarker may appear of practical interest, given the recent global spread of the *Aedes* mosquito vector.

In this biomarker, one may question the relevance of the presence of neutrophil-related genes, since the study was made on PBMC cells, in which only those traces of neutrophils remain that are not removed during centrifugation. The performance validation of our biomarker on whole blood samples confirms that neutrophil-related genes are not an artifact of bad purification of PBMC cells, but a real biological signal. Therefore, in future experiments, it may be of interest to keep neutrophils in samples if we want to precisely study their impact.

6.2.2 Methodological results

Interaction networks

From a methodological point of view, in Chapters 2 to 4, we explored network methods designed to find genes with relatively high scores of association to severe dengue and that interact. From Chapter 2, it appears that results obtained with different input networks may differ a lot. Therefore, improved curation of networks, trying several networks, and different sizes of input networks, in terms of nodes and connectivity, may be of great practical interest in future work.

Network search algorithms

The algorithm used for analysis also impacts strongly the results, as we see when comparing results from STRING interaction network in Chapters 2 and 4. Most remarkably, the biggest bottleneck is rather the interpretation of results, and their experimental validation. Statistical significance helps to gain statistical confidence in results, but most importantly generating small networks is key for a better interpretation and outcome.

Predictive models

An other approach to increase confidence in computational results may be to design predictive models, where we measure directly how well our results generalise, rather than aiming for statistical association in one specific dataset. This gives a better measure of how generalisable the results are. Nevertheless, one should pay attention again to the interpretability of the results, by, for instance limiting the number of features used in the resulting feature. The ensemble monotonic regression approach has the advantage of generating results where we control the number of features used, that are easy to interpret, and that we can visualize on a plot. In terms of performance, on dengue transcriptomic data it appears to generate more reproducible results than classical methods – even those that perform feature selec-

tion, such as logistic Lasso. This will probably not be the case for all datasets, since this regression depends on a monotonic model, which may not be the most appropriate choice for other datasets. Compared to a feature selection model such as Lasso, a key difference is that the ensemble aggregation method used here keeps all correlated pairs in the biomarker. This may be redundant in a non-noisy setting, but given the variability of expression measures in between individuals at a given time point, it may be interesting to keep redundant pairs in a gene expression signature. Moreover, the LARS algorithm used for Lasso picks genes iteratively, meaning that once the first gene picked, the choice of the second genes depends on this first gene [Efron et al., 2004]. This may make the actual set of chosen features highly variable based on small differences in the learning set.

However, ensemble monotonic regression is only one approach among others. We may consider, for instance, using support vector machine ensemble classifiers, as used in [Zak et al., 2016] to find a tuberculosis gene signature. For performance improvement, the conclusion of many challenges designed for method comparison (such as DREAM challenges) is that usually the best performance is obtained by "the wisdom of the crowd": These methods aggregate results of many different algorithms, by for instance, letting each method (or only several best methods) vote for the phenotype, and taking as outcome the majority vote [Marbach et al., 2012, Eduati et al., 2015]. Nevertheless, the disadvantage of such a method is that the features are then difficult to interpret, and usually many features are used.

In the case of an ensemble classifier, we may aggregate features otherwise than by simply using the most predictive pairs of features. To further make the model more robust and simpler, we may use, for instance, an approach that would remove a subset from the learning set, calculate a gene signature on the remaining set, iterate several times this procedure by leaving out different subsets and keeping for the final biomarker those genes that were appearing in all the computed gene signatures.

Another alternative to strengthen interpretability would be to first group genes in sets by common properties and try to apply machine learning methods to those groups taken

together. For instance, grouped Lasso is an adaptation of Lasso for such an approach [Yuan and Lin, 2006].

6.3 What if I had another three years for this project?

The next step of this project consists, in my opinion, in choosing the most interesting biological findings that can be validated in laboratories, and create experiments to validate them. This step will require a good knowledge of possible experiments on dengue; many of them may be hampered by the unavailability of an animal model that reproduces human severe symptoms after dengue infection. To be able to choose the most promising hypotheses to validate, we would likewise need to precisely understand the methods that have generated the statistically significant results.

That is why I would ask for financing for a common project between an experimental biologist and a computational biologist. Such a collaboration would also enable us to use the validation results to improve computational findings, upon which new experiments may then be executed using these results.

Moreover, I would invest time in better understanding the details of dengue immunology to be able, myself, to better choose among statistically significant results those that most echo with current knowledge of the disease.

In terms of methodology, ensemble monotonic regression needs to be thoroughly compared to other methods on a benchmark of diverse datasets for a precise assessment of the use of such a method. Creating a user-friendly implementation for instance in the Python machine-learning package sklearn may ease such an analysis.

Bibliography

[Akula et al., 2011] Akula, N., Baranova, A., Seto, D., Solka, J., Nalls, M. a., Singleton, A., Ferrucci, L., Tanaka, T., Bandinelli, S., Cho, Y. S., Kim, Y. J., Lee, J.-Y., Han, B.-G., and McMahon, F. J. (2011). A network-based approach to prioritize results from genome-wide association studies. *PloS one*, 6(9):e24220.

[Alcaraz et al., 2014] Alcaraz, N., Pauling, J., Batra, R., Barbosa, E., Junge, A., Christensen, A. G. L., Azevedo, V., Ditzel, H. J., and Baumbach, J. (2014). KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. *BMC systems biology*, 8(1):99.

[Alon, 2003] Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science (New York, N.Y.)*, 301(5641):1866–7.

[Amaratunga and Cabrera, 2001] Amaratunga, D. and Cabrera, J. (2001). Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association*, 96(456):1161–1170.

[Avirutnan et al., 2006] Avirutnan, P., Punyadee, N., Noisakran, S., Komoltri, C., Thiemmeca, S., Auethavornanan, K., Jairungsri, A., Kanlaya, R., Tangthawornchaikul, N., Puttikhunt, C., Pattanakitsakul, S.-n., Yenchitsomanus, P.-t., Mongkolsapaya, J., Kasinrerk, W., Sittisombut, N., Husmann, M., Blettner, M., Vasanawathana, S., Bhakdi, S., and Malasit, P. (2006). Vascular Leakage in Severe Dengue Virus Infections : A

Potential Role for the Nonstructural Viral Protein NS1 and Complement. *J Infect Dis*, 193:1078–88.

[Backes et al., 2012] Backes, C., Rurainski, A., Klau, G. W., Müller, O., Stöckel, D., Gerasch, A., Küntzer, J., Maisel, D., Ludwig, N., Hein, M., Keller, A., Burtscher, H., Kaufmann, M., Meese, E., and Lenhof, H.-P. (2012). An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic acids research*, 40(6):e43.

[Bakir-Gungor and Sezerman, 2011] Bakir-Gungor, B. and Sezerman, O. U. (2011). A new methodology to associate SNPs with human diseases according to their pathway related context. *PloS one*, 6(10):e26277.

[Bandyopadhyay et al., 2006] Bandyopadhyay, S., Lum, L. C. S., and Kroeger, A. (2006). Classifying dengue : a review of the difficulties in using the WHO case classification for dengue haemorrhagic fever. *Tropical Medicine and International Health*, 11(8):1238–1255.

[Barabási and Oltvai, 2004] Barabási, A.-l. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functionnal organisation. *Nature Reviews Genetics*, 5(February):101–113.

[Basuki et al., 2010] Basuki, P. S., Puspitasari, D., Husada, D., Darmowandowo, W., Soegijanto, S., and Yamanaka, A. (2010). Application of revised dengue classification criteria as a severity marker of dengue viral infection in Indonesia. *Southeast Asian J Trop Med Public Health*, 41(5):1088–1094.

[Batra et al., 2017] Batra, R., Alcaraz, N., Gitzhofer, K., Pauling, J., Ditzel, H. J., Hellmuth, M., Baumbach, J., and List, M. (2017). On the performance of de novo pathway enrichment. *npj Systems Biology and Applications*, 3(1):6.

[Beisser et al., 2010] Beisser, D., Klau, G. W., Dandekar, T., Müller, T., and Dittrich, M. T. (2010). BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics (Oxford, England)*, 26(8):1129–30.

[Biomarkers Definitions Working Group, 2001] Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology And Therapeutics*, 69(3):89–95.

[Biswas et al., 2015] Biswas, H. H., Gordon, A., Nunez, A., Perez, M. A., Balmaseda, A., and Harris, E. (2015). Lower Low-Density Lipoprotein Cholesterol Levels Are Associated with Severe Dengue Outcome. *PLoS Neglected Tropical Diseases*, 9(9):e0003904.

[Bourgon et al., 2010] Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *PNAS*, 107(21):9546–9551.

[Bravo et al., 1987] Bravo, J., Guzman, M., and Kouri, G. (1987). Why dengue haemorrhagic fever in Cuba? 1. Individual risk factors for dengue haemorrhagic fever/dengue shock syndrome (DHF/DSS). *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 81(5):816–20.

[Brown and Jurisica, 2007] Brown, K. R. and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*, 8(5):R95.

[Burrack and Morrison, 2014] Burrack, K. S. and Morrison, T. E. (2014). The role of myeloid cell activation and arginine metabolism in the pathogenesis of virus-induced diseases. *Frontiers in Immunology*, 5(AUG):428.

[Cattaert et al., 2011] Cattaert, T., Calle, M. L., Dudek, S. M., Mahachie John, J. M., Van Lishout, F., Urrea, V., Ritchie, M. D., and Van Steen, K. (2011). Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. *Annals of human genetics*, 75(1):78–89.

[Chuang et al., 2007] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(140):140.

[Churdboonchart et al., 1983] Churdboonchart, V., Bhamarapravati, N., and Futrakul, A. (1983). Crossed immunoelectrophoresis for the detection split products of the third complement in fever I. *Am. J. Trop. Med. Hyg.*, 32(3):569–76.

[Coffey et al., 2009] Coffey, L. L., Mertens, E., Brehin, A.-C., Fernandez-Garcia, M. D., Amara, A., Després, P., and Sakuntabhai, A. (2009). Human genetic determinants of dengue virus susceptibility. *Microbes and infection / Institut Pasteur*, 11(2):143–56.

[Coles, 2001] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

[Cookson et al., 2009] Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature reviews. Genetics*, 10(3):184–94.

[Croft et al., 2011] Croft, D., Kelly, G. O., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., Eustachio, P. D., and Stein, L. (2011). Reactome : a database of reactions , pathways and biological processes. *Nucleic Acids Research*, 39(Database issue):D691–D697.

[Dao et al., 2011] Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., and Sahinalp, S. C. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13):205–213.

[Deen et al., 2006] Deen, J. L., Harris, E., Wills, B., Balmaseda, A., Hammond, S. N., Rocha, C., Dung, N. M., and Rivera, M. D. J. (2006). The WHO dengue classification and case definitions : time for a reassessment. *Population (English Edition)*, 368:170–173.

[Dengue Vaccine Initiative, 2017] Dengue Vaccine Initiative (2017). Vaccine Development.

[Devignot et al., 2010a] Devignot, S., Sapet, C., Duong, V., Bergon, A., Rihet, P., Ong, S., Lorn, P. T., Chroeung, N., Ngeav, S., Tolou, H. J., Buchy, P., and Couissinier-Paris, P. (2010a). Genome-wide expression profiling deciphers host responses altered during

dengue shock syndrome and reveals the role of innate immunity in severe dengue. *PLoS ONE*, 5(7).

[Devignot et al., 2010b] Devignot, S., Sapet, C., Duong, V., Bergon, A., Rihet, P., Ong, S., Lorn, P. T., Chroeung, N., Ngeav, S., Tolou, H. J., Buchy, P., and Couissinier-Paris, P. (2010b). Genome-wide expression profiling deciphers host responses altered during dengue shock syndrome and reveals the role of innate immunity in severe dengue. *PLoS ONE*, 5(7):e11671.

[Dittrich et al., 2008] Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics (Oxford, England)*, 24(13):i223–31.

[Duong et al., 2015] Duong, V., Lambrechts, L., Paul, R. E., Ly, S., Lay, R. S., Long, K. C., Huy, R., Tarantola, A., Scott, T. W., Sakuntabhai, A., and Buchy, P. (2015). Asymptomatic humans transmit dengue virus to mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):14688–93.

[Edfors et al., 2016] Edfors, F., Danielsson, F., Hallström, B. M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., and Uhlén, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. pages 1–10.

[Eduati et al., 2015] Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., Norman, T., Kellen, M., Menden, M. P., Yang, J., Zhan, X., Zhong, R., Xiao, G., Xia, M., Abdo, N., Kosyk, O., Dream, N.-n.-u., Collaboration, T., Friend, S., Dearry, A., Simeonov, A., Tice, R. R., Rusyn, I., and Wright, F. A. (2015). A n a ly s i s Prediction of human population responses to toxic compounds by a collaborative competition. *Nature biotechnology*, 33(9):933–940.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). LEAST ANGLE REGRESSION 1 . Introduction . Automatic model-building algorithms are familiar , and sometimes notorious , in the linear model literature : Forward Selection ,

Backward Elimination , All Subsets regression and various combinations are used to. *The Annals of Statistics*, 32(2):407–499.

[Efron and Tibshirani, 1997] Efron, B. and Tibshirani, R. (1997). Improvements on Cross-Validation: The .632 Bootstrap Method. *Journal of the American Statistical Association*, 92(438):548–560.

[Eichler et al., 2010] Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Publishing Group*, 11(6):446–450.

[Elgueta et al., 2009] Elgueta, R., Benson, M. J., de Vries, V. C., Wasiuk, A., Guo, Y., and Noelle, R. J. (2009). Molecular mechanism and function of CD40/CD40L engagement in the immune system. *Immunol Rev*, 229(1):189.

[Fabregat et al., 2016] Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., Mckay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and Eustachio, P. D. (2016). The Reactome pathway Knowledgebase. *Nuclei Acids Research*, 44(Database issue):481–487.

[Fontana et al., 2015] Fontana, M. F., Baccarella, A., Pancholi, N., Pufall, A., Herbert, D. B. R., and Kim, C. C. (2015). JUNB Is a Key Transcriptional Modulator of Macrophage Activation. *The Journal of Immunology*, 194(1):177–186.

[Franceschini et al., 2013] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue):D808–15.

[Francis, 1960] Francis, T. (1960). On the Doctrine of Original Antigenic Sin. *Proceedings of the American Philosophical Society*, 104(6):572–578.

[Gamazon et al., 2010] Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner,

E. O., Nicolae, D. L., Dolan, M. E., and Cox, N. J. (2010). SCAN: SNP and copy number annotation. *Bioinformatics*, 26(2):259–262.

[Gandini et al., 2011] Gandini, M., Reis, S. R. N. I., Torrentes-Carvalho, A., Azeredo, E. L., da Silva Freire, M., Galler, R., and Kubelka, C. F. (2011). Dengue-2 and yellow fever 17DD viruses infect human dendritic cells, resulting in an induction of activation markers, cytokines and chemokines and secretion of different TNF-alpha and IFN-alpha profiles. *Memorias do Instituto Oswaldo Cruz*, 106(5):594–605.

[García-Alonso et al., 2012] García-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., and Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic acids research*, 40(20):e158.

[Grange, 2014] Grange, L. (2014). *Epistasis in genetic susceptibility to infectious diseases. Comparison and Development of Methods. Application to Severe Dengue in Asia*. PhD thesis, Paris Diderot.

[Grundberg et al., 2012] Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Keildson, S., Bell, J. T., Yang, T.-p., Meduri, E., Nisbett, J., Sekowska, M., Wilk, A., Shin, S.-y., Travers, M., Min, J. L., Ring, S., Ho, K., Thorleifsson, G., Kong, A., Thorsteindottir, U., Ainali, C., Dimas, A. S., Zondervan, K. T., Ahmadi, K. R., Schadt, E. E., Stefansson, K., and Spector, T. D. (2012). Mapping cis - and trans -regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089.

[Guzman and Kouri, 2002] Guzman, M. and Kouri, G. (2002). Dengue: an update. *Lancet Infectious Diseases*, 2(1):33–42.

[Guzman et al., 2010] Guzman, M. G., Halstead, S. B., Artsob, H., Buchy, P., Farrar, J., Gubler, D. J., Hunsperger, E., Kroeger, A., Margolis, H. S., Martínez, E., Nathan, M. B., Pelegrino, J. L., Simmons, C., Yoksan, S., and Peeling, R. W. (2010). Dengue: a continuing global threat. *Nature reviews. Microbiology*, 8(12 Suppl):S7–16.

[Gwinner et al., 2016] Gwinner, F., Boulday, G., Vandiedonck, C., Arnould, M., Cardoso, C., Nikolayeva, I., Guitart-Pla, O., Denis, C. V., Christophe, O. D., Beghain, J., Tournier-Lasserve, E., and Schwikowski, B. (2016). Network-based analysis of omics data: The LEAN method. *Bioinformatics (Oxford, England)*, 33(5)(December 2016):701–709.

[Hadinegoro et al., 2015] Hadinegoro, S., Arredondo-García, J., Capeding, M., Deseda, M., Chotpitayasunondh, T., Dietze, R., Muhammad Ismail, H. H., Reynales, H., Limkittikul, K., Rivera-Medina, D., Tran, H., Bouckenooghe, D. Chansinghakul, M. Cortés, K. Fanouillere, R. Forrat, C. Frago, S. Gailhardou, N. Jackson, F. Noriega, E. Plennevaux, T.A. Wartel, B. Zambrano, and M. Saville, f. t. C.-T. D. V. W. G., and Abstract (2015). Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. *The New England journal of medicine*, 373(13):1195–1206.

[Hadinegoro, 2012] Hadinegoro, S. R. S. (2012). The revised WHO dengue case classification : does the system need to be modified ? *Paediatrics and International Child Health*, 32(Suppl 1):33–38.

[Halstead, 2003] Halstead, S. B. (2003). Neutralisation and antibody-dependent enhancement of dengue viruses. In *Advances in virus research*, volume 60, pages 421–467.

[Halstead, 2014] Halstead, S. B. (2014). Dengue Antibody-Dependent Enhancement : Knowns and Unknowns. *Microbiology Spectrum*, 2(6):AID–0022–2014.

[Halstead et al., 2001] Halstead, S. B., Streit, T. G., Lafontant, J. G. U. Y., Putvatana, R., Russell, K., Sun, W., Kanesa-thasan, N., Hayes, C. G., and Watts, D. M. (2001). Haiti: absence of dengue hemorrhagic fever despite hyperendemic dengue virus transmission. *Am J Trop Med Hyg.*, 65(3):180–183.

[Hartwell et al., 1999] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements*

of Statistical Learning: : Data Mining, Inference, and Prediction. Springer Series in Statistics, second edi edition.

[Hecker et al., 2017] Hecker, J., Maaser, A., Prokopenko, D., Fier, H. L., and Lange, C. (2017). Reporting Correct p Values in VEGAS Analyses. *Twin Research and Human Genetics*, 20(3):257–259.

[Hiersche et al., 2013] Hiersche, M., Rühle, F., and Stoll, M. (2013). Postgwas: advanced GWAS interpretation in R. *PloS one*, 8(8):e71775.

[Hinkes et al., 2006] Hinkes, B., Wiggins, R. C., Gbadegesin, R., Vlangos, C. N., Seelow, D., Nürnberg, G., Garg, P., Verma, R., Chaib, H., Hoskins, B. E., Ashraf, S., Becker, C., Hennies, H. C., Goyal, M., Wharram, B. L., Schachter, A. D., Mudumana, S., Drummond, I., Kerjaschki, D., Waldherr, R., Dietrich, A., Ozaltin, F., Bakkaloglu, A., Cleper, R., Basel-Vanagaite, L., Pohl, M., Griebel, M., Tsygin, A. N., Soylu, A., Müller, D., Sorli, C. S., Bunney, T. D., Katan, M., Liu, J., Attanasio, M., O'toole, J. F., Hasselbacher, K., Mucha, B., Otto, E. a., Airik, R., Kispert, A., Kelley, G. G., Smrcka, A. V., Gudermann, T., Holzman, L. B., Nürnberg, P., and Hildebrandt, F. (2006). Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nature genetics*, 38(12):1397–405.

[Hoang et al., 2010] Hoang, L. T., Lynn, D. J., Henn, M., Birren, B. W., Lennon, N. J., Le, P. T., Duong, K. T. H., Nguyen, T. T. H., Mai, L. N., Farrar, J. J., Hibberd, M. L., and Simmons, C. P. (2010). The early whole-blood transcriptional signature of dengue virus and features associated with progression to dengue shock syndrome in Vietnamese children and young adults. *Journal of virology*, 84(24):12982–12994.

[Hormozdiari et al., 2015] Hormozdiari, F., Penn, O., Borenstein, E., and Eichler, E. E. (2015). The discovery of integrated gene networks for autism and related disorders.

[Ideker et al., 2002] Ideker, T., Ozier, O., Schwikowski, B., and Andrew, F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl.:233–240.

[Iorio et al., 2016] Iorio, F., Knijnenburg, T. A., Vis, D. J., Saez-rodriguez, J., Mcdermott, U., and Garnett, M. J. (2016). Resource A Landscape of Pharmacogenomic Interactions in Resource A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754.

[Janeway CA et al., 2001] Janeway CA, J., Travers, P., Walport, M., and Al., E. (2001). The major histocompatibility complex and its functions. In Garland Science, editor, *Immunobiology: The Immune System in Health and Disease. 5th edition.* New York, New York, USA.

[Jia and Zhao, 2013] Jia, P. and Zhao, Z. (2013). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human genetics*, (Gibson 2011).

[Jia et al., 2011] Jia, P., Zheng, S., Long, J., Zheng, W., and Zhao, Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1):95–102.

[Jin et al., 2008] Jin, G., Zhou, X., Wang, H., Zhao, H., Cui, K., Zhang, X. S., Chen, L., Hazen, S. L., Li, K., and Wong, S. T. C. (2008). The knowledge-integrated network biomarkers discovery for major adverse cardiac events. *Journal of Proteome Research*, 7(9):4013–4021.

[John et al., 2015] John, D. V., Lin, Y.-S., and Perng, G. C. (2015). Biomarkers of severe dengue disease - a review. *Journal of biomedical science*, 22:83.

[Kamburov et al., 2011] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic acids research*, 39(Database issue):D712–7.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes. *Oxford University Press*, 28(1):27–30.

[Karlin and Altschul, 1990] Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring

schemes. *Proceedings of the National Academy of Sciences of the United States of America*, 87(March):2264–2268.

[Karolchik et al., 2004] Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nuclei Acids Research*, 32(9):493–496.

[Keshava Prasad et al., 2009] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. I., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database - 2009 update. *Nucleic Acids Research*, 37(SUPPL. 1):767–772.

[Khatri et al., 2012] Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2).

[Khor et al., 2011] Khor, C. C., Chau, T. N. B., Pang, J., Davila, S., Long, H. T., Ong, R. T. H., Dunstan, S. J., Wills, B., Farrar, J., Van Tram, T., Gan, T. T., Binh, N. T. N., Tri, L. T., Lien, L. B., Tuan, N. M., Tham, N. T. H., Lanh, M. N., Nguyet, N. M., Hieu, N. T., Van N Vinh Chau, N., Thuy, T. T., Tan, D. E. K., Sakuntabhai, A., Teo, Y.-Y., Hibberd, M. L., and Simmons, C. P. (2011). Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nature Genetics*, 43(11):1139–1141.

[Kohavi and Provost, 1998] Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine learning*, 30:271–274.

[Kwissa et al., 2014] Kwissa, M., Nakaya, H. I., Onlamoon, N., Wrammert, J., Villinger, F., Perng, G. C., Yoksan, S., Pattanapanyasat, K., Chokephaibulkit, K., Ahmed, R.,

and Pulendran, B. (2014). Resource Dengue Virus Infection Induces Expansion that Stimulates Plasmablast Differentiation. *Cell Host and Microbe*, 16(1):115–127.

[Lan et al., 2008] Lan, N. T. P., Kikuchi, M., Huong, V. T. Q., Ha, D. Q., Thuy, T. T., Tham, D., Tuan, H. M., Tuong, V. V., Nga, C. T. P., Dat, T. V., Oyama, T., Morita, K., Yasunami, M., and Hirayama, K. (2008). Protective and Enhancing HLA Alleles , HLA-DRB1 * 0901 and HLA-A * 24 , for Severe Forms of Dengue Virus Infection , Dengue Hemorrhagic Fever and Dengue Shock Syndrome. *PLoS Negl Trop Dis*, 2(10):e304.

[Lee et al., 2011] Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research*, 21(7):1109–21.

[Lee et al., 2016] Lee, I.-K., Liu, J.-W., Chen, Y.-H., Chen, Y.-C., Tsai, C.-Y., Huang, S.-Y., Lin, C.-Y., and Huang, C.-H. (2016). Development of a Simple Clinical Risk Score for Early Prediction of Severe Dengue in Adult Patients. *Plos One*, 11(5):e0154772.

[Liang et al., 2013] Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Abecasis, G. R., Moffatt, M. F., and Cookson, W. O. C. (2013). A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome research*, 23(4):716–26.

[Lin et al., 2008] Lin, C.-f., Wan, S.-w., Chen, M.-c., Lin, S.-c., Cheng, C.-c., Chiu, S.-c., Hsiao, Y.-l., Lei, H.-y., Liu, H.-s., Yeh, T.-m., and Lin, Y.-s. (2008). Liver injury caused by antibodies against dengue virus nonstructural protein 1 in a murine model. *Laboratory Investigation*, 88(December 2007):1079–1089.

[Liu et al., 2010] Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *American journal of human genetics*, 87(1):139–45.

[Liu et al., 2007] Liu, M., Liberzon, A., Sek, W. K., Lai, W. R., Park, P. J., Kohane, I. S.,

and Kasif, S. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genetics*, 3(6):0958–0972.

[Liu et al., 2017] Liu, Y., Brossard, M., Roqueiro, D., Margaritte-Jeannin, P., Sarnowski, C., Bouzigon, E., and Demenais, F. (2017). SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. *Bioinformatics*, (January):btx004.

[Long et al., 2009] Long, H. T., Hibberd, M. L., Hien, T. T., Dung, N. M., Van Ngoc, T., Farrar, J., Wills, B., and Simmons, C. P. (2009). Patterns of gene transcript abundance in the blood of children with severe or uncomplicated dengue highlight differences in disease evolution and host response to dengue virus infection. *The Journal of infectious diseases*, 199(4):537–46.

[Lonsdale et al., 2013] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

[Luft, 1965] Luft, J. H. (1965). The ultrastructural basis of vascular permeability. In Zweifach, B. W., Grant, L., and McCluskey, R. T., editors, *The inflammatory process*. Academic Press, Hey York.

[Maere et al., 2005] Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, 21(16):3448–9.

[Maher, 2008] Maher, B. (2008). The case of the missing heritability. *Nature*, 456(November):18–21.

[Marbach et al., 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Consortium, T. D., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804.

[Martina et al., 2009] Martina, B. E. E., Koraka, P., and Osterhaus, A. D. M. E. (2009). Dengue virus pathogenesis: An integrated view. *Clinical Microbiology Reviews*, 22(4):564–581.

[Merico et al., 2010] Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (2010). Enrichment Map : A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *Plos One*, 5(11):e13984.

[Mishra and MacGregor, 2017] Mishra, A. and MacGregor, S. (2017). A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Research and Human Genetics*, 20(1):1–9.

[Mishra and Macgregor, 2017] Mishra, A. and Macgregor, S. (2017). VEGAS2 : Software for More Flexible Gene-Based Testing. *Twin Research and Human Genetics*, 18(1):86–91.

[Mitra et al., 2013] Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732.

[Munder et al., 2005] Munder, M., Mollinedo, F., Calafat, J., Canchado, J., Gil-lamaignere, C., Fuentes, M., Luckner, C., Doschko, G., Eichmann, K., Mu, F.-m., Ho, A. D., Goerner,

M., and Modolell, M. (2005). Arginase I is constitutively expressed in human granulocytes and participates in fungicidal activity. *Immunobiology*, 105(6):2549–2556.

[Nacu et al., 2007] Nacu, S., Critchley-Thorne, R., Lee, P., and Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics (Oxford, England)*, 23(7):850–8.

[Narvaez et al., 2011] Narvaez, F., Gutierrez, G., Perez, M. A., Elizondo, D., Nunez, A., Balmaseda, A., and Harris, E. (2011). Evaluation of the traditional and revised WHO classifications of dengue disease severity. *PLoS Neglected Tropical Diseases*, 5(11):1–8.

[Nhi et al., 2016] Nhi, D. M., Huy, N. T., Ohyama, K., Kimura, D., Thi, N., Lan, P., Uchida, L., Thuong, N. V., Thi, C., Nhon, M., Phuc, L. H., Mai, N. T., Mizukami, S., Bao, L. Q., Doan, N. N., Binh, N. V. T., Quang, L. C., Karbwang, J., Yui, K., Morita, K., Thi, V., Huong, Q., and Hirayama, K. (2016). A Proteomic Approach Identifies Candidate Early Biomarkers to Predict Severe Dengue in. *PLoS Neglected Tropical Diseases*, 10(2):e0004435.

[Nicolae et al., 2010] Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888.

[Normile, 2013] Normile, D. (2013). Surprising New Dengue Virus Throws A Spanner in Disease Control Efforts. *Science*, 342(6157):415.

[Olex et al., 2014] Olex, A. L., Turkett, W. H., Fetrow, J. S., and Loeser, R. F. (2014). Integration of gene expression data with network-based analysis to identify signaling and metabolic pathways regulated during the development of osteoarthritis. *Gene*, 542(1):38–45.

[Pang et al., 2016] Pang, J., Lindblom, A., Tolfvenstam, T., Thein, T. L., Naim, A. N. M., Ling, L., Chow, A., Chen, M. I. C., Ooi, E. E., Leo, Y. S., and Hibberd, M. L. (2016).

Discovery and validation of prognostic biomarker models to guide triage among adult dengue patients at early infection. *PLoS ONE*, 11(6):1–19.

[Pedregosa et al., 2012] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pérez-Palma et al., 2016] Pérez-Palma, E., Andrade, V., Caracci, M. O., Bustos, B. I., Villaman, C., Medina, M. A., Ávila, M. E., Ugarte, G. D., and De Ferrari, G. V. (2016). Early Transcriptional Changes Induced by Wnt/ β -Catenin Signaling in Hippocampal Neurons. *Neural Plasticity*, 2016:1–13.

[Peri et al., 2003] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., and Chakravarti, A. (2003). Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research*, 13:2363–2371.

[Phuong et al., 2004] Phuong, C. X. T., Nhan, N. T., Kneen, R., Thuy, P. T. T., Thien, C. H. U. V. A. N., Nga, T., Thuy, T. T., Solomon, T., Stepniewska, K., Wills, B., and The Dong Nai Study Group (2004). Clinical diagnosis and assessment of severity of confirmed dengue infections in Vietnamese children: is the world health organization classification system helpful? *The American Society of Tropical Medicine and Hygiene*, 70(2):172–179.

[Popper et al., 2012] Popper, S. J., Gordon, A., Liu, M., Balmaseda, A., Harris, E., and Relman, D. A. (2012). Temporal dynamics of the transcriptional response to dengue virus infection in Nicaraguan children. *PLoS neglected tropical diseases*, 6(12):e1966.

[Priyamvada et al., 2016] Priyamvada, L., Quicke, K. M., Hudson, W. H., Onlamoon, N., and Sewatanon, J. (2016). Human antibody responses after dengue virus infection are highly cross-reactive to Zika virus. *PNAS*, 113(28):7852–7857.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75.

[Qiu et al., 2010] Qiu, Y.-Q., Zhang, S., Zhang, X.-S., and Chen, L. (2010). Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC bioinformatics*, 11:26.

[Rajagopalan and Agarwal, 2005] Rajagopalan, D. and Agarwal, P. (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics (Oxford, England)*, 21(6):788–93.

[Ratajczak-Wrona et al., 2012] Ratajczak-Wrona, W., Jablonska, E., Garley, M., Jablonski, J., Radziwon, P., and Iwaniuk, A. (2012). Role of AP-1 family proteins in regulation of inducible nitric oxide synthase (iNOS) in human neutrophils. *Journal of immunotoxicology*, 10(1):32–39.

[Rodenhuis-Zybert et al., 2010] Rodenhuis-Zybert, I. A., Wilschut, J., and Smit, J. M. (2010). Dengue virus life cycle : viral and host factors modulating infectivity. *Cell. Mol. Life Sci.*, (67):2773–2786.

[Rothman, 2011] Rothman, A. L. (2011). Immunity to dengue virus: a tale of original antigenic sin and tropical cytokine storms. *Nature reviews. Immunology*, 11(8):532–543.

[Sangkawibha et al., 1984] Sangkawibha, N., Rojanasuphot, S., Ahandrik, S., Viriyapongse,

S., Jatanasen, S., Salitul, V., Phanthumachinda, B., and Halstead, S. (1984). Risk factors in dengue shock syndrome: a prospective epidemiologic study in Rayong, Thailand. I. The 1980 outbreak. *American Journal of Epidemiology*, 120(5):653–69.

[Schadt et al., 2008] Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., Johnson, J. M., Rohl, C. A., Van Nas, A., Mehrabian, M., Drake, T. A., Lusis, A. J., Smith, R. C., Guengerich, F. P., Strom, S. C., Schuetz, E., Rushmore, T. H., and Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, 6(5):1020–1032.

[Scitable by Nature Education, 2014] Scitable by Nature Education (2014). "single-nucleotide polymorphism / SNP".

[Screaton et al., 2015] Screaton, G., Mongkolsapaya, J., Yacoub, S., and Roberts, C. (2015). New insights into the immunopathology and control of dengue virus infection. *Nature Reviews Immunology*, 15(12):745–759.

[Shaio et al., 1992] Shaio, M.-f., Chang, F.-y., and Hou, S.-c. (1992). Complement pathway activity in serum from patients with classical dengue fever. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 86:672–675.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 11:2498–2504.

[Sharma et al., 2013] Sharma, A., Gulbahce, N., Pevzner, S. J., Menche, J., Ladenvall, C., Folkersen, L., Eriksson, P., Orho-Melander, M., and Barabási, A.-L. (2013). Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits.

[Simmons et al., 2012] Simmons, C. P., Farrar, J. J., Van Vinh Chau, N., and Wills, B. (2012). Dengue current concepts. *The New England journal of medicine*, 366(15):1423–1432.

[Simmons et al., 2007a] Simmons, C. P., Popper, S., Dolocek, C., Chau, T. N. B., Griffiths, M., Dung, N. T. P., Long, T. H., Hoang, D. M., Chau, N. V., Thao, L. T. T., Hien, T. T., Relman, D. A., and Farrar, J. (2007a). Patterns of Host GenomeWide Gene Transcript Abundance in the Peripheral Blood of Patients with Acute Dengue Hemorrhagic Fever. *Journal of Infectious Diseases*, 195(8):1097–1107.

[Simmons et al., 2007b] Simmons, C. P., Popper, S., Dolocek, C., Chau, T. N. B., Griffiths, M., Dung, N. T. P., Long, T. H., Hoang, D. M., Chau, N. V., Thao, L. T. T., Hien, T. T., Relman, D. A., and Farrar, J. (2007b). Patterns of Host GenomeWide Gene Transcript Abundance in the Peripheral Blood of Patients with Acute Dengue Hemorrhagic Fever. *Journal of Infectious Diseases*, 195(8):1097–1107.

[Slatkin, 2008] Slatkin, M. (2008). Linkage disequilibrium understanding the evolutionary past and mapping the medical future. *Nature reviews Genetics*, 9(6):477–485.

[Smith et al., 2014] Smith, C. L., Dickinson, P., Forster, T., Craigon, M., Ross, A., Khondoker, M. R., France, R., Ivens, A., Lynn, D. J., Orme, J., Jackson, A., Lacaze, P., Flanagan, K. L., Stenson, B. J., and Ghazal, P. (2014). Identification of a human neonatal immune-metabolic network associated with bacterial infection. *Nature communications*, 5:4649.

[Soundravally et al., 2015] Soundravally, R., Agieshkumar, B., Daisy, M., Sherin, J., and Cleetus, C. C. (2015). Ferritin levels predict severe dengue. *Infection*, 43:13–19.

[Stephens et al., 2002] Stephens, H., Klaythong, R., Sirikong, M., Vaughn, D. W., Green, S., Kalayanarooj, S., Endy, T. P., Libraty, D. H., Nisalak, A., Innis, B. L., Rothman, A. L., Ennis, F. A., and Chandanayingyong, D. (2002). HLA-A and -B allele associations with secondary dengue virus infections correlate with disease severity and the infecting viral serotype in ethnic Thais. *Tissue Antigens*, 60(1):309–318.

[Stettler et al., 2016] Stettler, K., Beltramello, M., Espinosa, D. A., Graham, V., Cassotta, A., Bianchi, S., Vanzetta, F., Minola, A., Jaconi, S., Mele, F., Foglierini, M., Pedotti, M., Simonelli, L., Dowall, S., Atkinson, B., Percivalle, E., Simmons, C. P., Varani, L., Blum, J., Baldanti, F., Cameroni, E., Hewson, R., Harris, E., Lanzavecchia, A., Sallusto, F., and Corti, D. (2016). Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. *Science*, 353(6301):823–827.

[Stouffer et al., 1949] Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams, R. M. (1949). The American soldier: Adjustment during Army life. (Studies in social psychology in World War II, Vol. I). *Princeton University Press*, 28(1):87–90.

[Stout, 2012] Stout, Q. F. (2012). Isotonic Regression via Partitioning. *Algorithmica*, 66(1):93–112.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., and Ebert, B. L. (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.

[Sun et al., 2016] Sun, P., Celluzzi, C. M., Marovich, M., Subramanian, H., Eller, M., Widjaja, S., Palmer, D., Porter, K., Sun, W., Sun, P., Celluzzi, C. M., Marovich, M., Subramanian, H., Eller, M., Widjaja, S., Palmer, D., Porter, K., Sun, W., and Burgess, T. (2016). CD40 Ligand Enhances Dengue Viral Infection of Dendritic Cells : A Possible Mechanism for T Cell-Mediated. *The Journal of Immunology*, 177:6497–6503.

[Szklarczyk et al., 2014] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2014). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(October 2014):447–452.

[Tezak et al., 2010] Tezak, Z., Kondratovitch, M. V., and Mansfield, E. (2010). US FDA and personalized medicine: in vitro diagnostic regulatory perspective. *Personalized Medicine*, 7(5):517–530.

[Thanachartwet et al., 2015] Thanachartwet, V., Oer-areemitr, N., Chamnanchanunt, S., Sahassananda, D., Jittmittraphap, A., Suwannakudt, P., Desakorn, V., and Wattanathum, A. (2015). Identification of clinical factors associated with severe dengue among Thai adults: a prospective study. *BMC Infectious Diseases*, 15(1):420.

[Tuan et al., 2017] Tuan, N. M., Nhan, H. T., Chau, N. V. V., Hung, N. T., and Manh, H. (2017). An evidence-based algorithm for early prognosis of severe dengue in the outpatient setting. *Clinical Infectious Diseases*, 64(5):656–63.

[Ubol et al., 2008] Ubol, S., Masrinoul, P., Chaijaruwanich, J., Kalayanarooj, S., Charoensirisuthikul, T., and Kasisith, J. (2008). Differences in global gene expression in peripheral blood mononuclear cells indicate a significant role of the innate responses in progression of dengue fever but not dengue hemorrhagic fever. *The Journal of infectious diseases*, 197(10):1459–67.

[Veyrieras et al., 2008] Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics*, 4(10).

[Wang et al., 2010] Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature reviews. Genetics*, 11(12):843–54.

[Wang et al., 2015a] Wang, L., Matsushita, T., Madireddy, L., Mousavi, P., and Baranzini, S. E. (2015a). PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics*, 31(2):262–264.

[Wang et al., 2015b] Wang, Q., Yu, H., Zhao, Z., and Jia, P. (2015b). EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics*, 31(March):2591–2594.

[West et al., 2013] West, J., Beck, S., Wang, X., and Teschendorff, A. E. (2013). An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Scientific reports*, 3:1630.

[Westra et al., 2013] Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., 't Hoen, P. A. C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Nalls, M. A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S. A., Enquobahrie, D. A., Lumley, T., Montgomery, G. W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R. C., Visscher, P. M., Knight, J. C., Psaty, B. M., Ripatti, S., Teumer, A., Frayling, T. M., Metspalu, A., van Meurs, J. B. J., and Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*, 45(10):1238–1243.

[Whitehorn et al., 2013] Whitehorn, J., Chau, T. N. B., Nguyet, N. M., Kien, D. T. H., Quyen, N. T. H., Trung, D. T., Pang, J., Wills, B., Van Vinh Chau, N., Farrar, J., Hibberd, M. L., Khor, C. C., and Simmons, C. P. (2013). Genetic variants of MICB and PLCE1 and associations with non-severe dengue. *PloS one*, 8(3):e59067.

[WHO, 2017] WHO (2017). "Dengue And Severe Dengue".

[WHO (World Health Organisation), 1997] WHO (World Health Organisation) (1997). Dengue haemorrhagic fever: diagnosis, treatment, prevention and control. 2nd edition. *Geneva : World Health Organization.*, 40(1):103–117.

[WHO (World Health Organisation), 2009] WHO (World Health Organisation) (2009). Dengue Guidelines for diagnosis, treatment, prevention and control.

[Xia et al., 2012] Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y. H., Wang, W., Zou, F., Sun, W., Sullivan, P. F., and Wright, F. A. (2012). SeeQTL: A searchable database for human eQTLs. *Bioinformatics*, 28(3):451–452.

[Yauch et al., 2010] Yauch, L. E., Prestwood, T. R., May, M. M., Morar, M. M., Zellweger, R. M., Peters, B., Sette, A., and Shresta, S. (2010). CD4+ T cells are not required

for the induction of dengue virus-specific CD8+ T cell or antibody responses but contribute to protection after vaccination. *Journal of immunology (Baltimore, Md. : 1950)*, 185(9):5405–16.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *ournal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67.

[Zak et al., 2016] Zak, D. E., Penn-Nicholson, A., Scriba, T. J., Thompson, E., Suliman, S., Amon, L. M., Mahomed, H., Erasmus, M., Whatney, W., Hussey, G. D., Abrahams, D., Kafaar, F., Hawkridge, T., Verver, S., Hughes, E. J., Ota, M., Sutherland, J., Howe, R., Dockrell, H. M., Boom, W. H., Thiel, B., Ottenhoff, T. H. M., Mayanja-Kizza, H., Crampin, A. C., Downing, K., Hatherill, M., Valvo, J., Shankar, S., Parida, S. K., Kaufmann, S. H. E., Walzl, G., Aderem, A., and Hanekom, W. A. (2016). A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet*, 6736(15):1–11.

Appendix

Appendix A

LEAN result: List of centers of significant stars

List of centers of significant stars: For each star center, are presented its gene symbol, Ensembl protein ID, the number of genes included in the significant star k, the total number of its neighbors m, and the statistical significance of the star score measured by a q-value.

Star_center	Star_center_ENSP	k	m	q-value
CD4	ENSP00000011653	112	226	0.009627943
TNFRSF1A	ENSP00000162749	53	95	0.009627943
NFKB2	ENSP00000189444	20	31	0.009627943
CDC6	ENSP00000209728	121	261	0.009627943
AAAS	ENSP00000209873	36	59	0.009627943
NDUFB7	ENSP00000215565	46	64	0.009627943
USP18	ENSP00000215794	27	34	0.009627943
SNRPD3	ENSP00000215829	91	142	0.009627943
RBX1	ENSP00000216225	79	160	0.009627943
NFKBIA	ENSP00000216797	74	119	0.009627943

RELB	ENSP00000221452	14	24	0.009627943
POLR2I	ENSP00000221859	99	169	0.009627943
OGDH	ENSP00000222673	32	41	0.009627943
PSMA2	ENSP00000223321	89	146	0.009627943
NUP107	ENSP00000229179	80	127	0.009627943
GMNN	ENSP00000230056	92	171	0.009627943
NUP155	ENSP00000231498	50	84	0.009627943
EIF2AK2	ENSP00000233057	23	51	0.009627943
NDUFS7	ENSP00000233627	36	50	0.009627943
EGR1	ENSP00000239938	52	106	0.009627943
CDKN1A	ENSP00000244741	93	247	0.009627943
IRF1	ENSP00000245414	53	82	0.009627943
NUP37	ENSP00000251074	87	164	0.009627943
DHX58	ENSP00000251642	17	25	0.009627943
DPF2	ENSP00000252268	11	11	0.009627943
NUP210	ENSP00000254508	35	57	0.009627943
OASL	ENSP00000257570	41	50	0.009627943
MPHOSPH6	ENSP00000258169	20	26	0.009627943
SP110	ENSP00000258381	15	17	0.009627943
RTP4	ENSP00000259030	30	35	0.009627943
NDUFB5	ENSP00000259037	49	70	0.009627943
RIPK1	ENSP00000259808	42	71	0.009627943
NUP133	ENSP00000261396	73	117	0.009627943
NUP153	ENSP00000262077	49	86	0.009627943
PARP12	ENSP00000263549	15	17	0.009627943
IFIH1	ENSP00000263642	36	44	0.009627943
NDUFS3	ENSP00000263774	48	81	0.009627943
HERC6	ENSP00000264346	20	23	0.009627943
HERC5	ENSP00000264350	27	41	0.009627943

NUP54	ENSP00000264883	37	61	0.009627943
EXOSC7	ENSP00000265564	19	25	0.009627943
GTF2H1	ENSP00000265963	43	76	0.009627943
SNRPF	ENSP00000266735	88	146	0.009627943
IRF8	ENSP00000268638	30	55	0.009627943
SNRPG	ENSP00000272348	95	153	0.009627943
IL18	ENSP00000280357	34	68	0.009627943
ERCC3	ENSP00000285398	42	74	0.009627943
UBE2L6	ENSP00000287156	21	41	0.009627943
U2AF1	ENSP00000291552	64	121	0.009627943
NUP35	ENSP00000295119	41	59	0.009627943
DTX3L	ENSP00000296161	10	13	0.009627943
POLR2H	ENSP00000296223	122	206	0.009627943
IFI27	ENSP00000298902	22	33	0.009627943
NDUFB8	ENSP00000299166	48	68	0.009627943
NOD2	ENSP00000300589	17	26	0.009627943
UBE2E1	ENSP00000303709	37	86	0.009627943
CDC40	ENSP00000304370	58	134	0.009627943
UBB	ENSP00000304697	127	256	0.009627943
NKRF	ENSP00000304803	17	19	0.009627943
ISG20	ENSP00000306565	20	28	0.009627943
LSM1	ENSP00000310596	31	46	0.009627943
DMAP1	ENSP00000312697	15	26	0.009627943
POLR2A	ENSP00000314949	127	220	0.009627943
NDUFS8	ENSP00000315774	41	59	0.009627943
CIITA	ENSP00000316328	29	55	0.009627943
EPSTI1	ENSP00000318982	23	24	0.009627943
NDUFV1	ENSP00000322450	42	57	0.009627943
EXOSC3	ENSP00000323046	19	25	0.009627943

POLR2L	ENSP00000324124	118	200	0.009627943
SAMD9L	ENSP00000326247	20	21	0.009627943
NCBP2	ENSP00000326806	110	229	0.009627943
IRF7	ENSP00000329411	48	84	0.009627943
MT1H	ENSP00000330587	6	7	0.009627943
IP6K2	ENSP00000331103	20	32	0.009627943
MX2	ENSP00000333657	26	34	0.009627943
SF3B1	ENSP00000335321	51	115	0.009627943
GBP5	ENSP00000340396	15	31	0.009627943
NUP43	ENSP00000342262	74	117	0.009627943
SNRPD2	ENSP00000342374	95	151	0.009627943
IFI6	ENSP00000342513	28	36	0.009627943
UBC	ENSP00000344818	506	1203	0.009627943
NUP50	ENSP00000345895	37	59	0.009627943
IRF5	ENSP00000349770	44	53	0.009627943
PARP9	ENSP00000353512	28	30	0.009627943
STAT1	ENSP00000354394	84	201	0.009627943
XAF1	ENSP00000354822	41	53	0.009627943
H3F3A	ENSP00000355778	57	128	0.009627943
IRF6	ENSP00000355988	26	38	0.009627943
UCHL5	ENSP00000356425	50	86	0.009627943
RNASEL	ENSP00000356530	27	30	0.009627943
ADAR	ENSP00000357459	15	32	0.009627943
BUB3	ENSP00000357858	57	152	0.009627943
MAP3K7	ENSP00000358335	59	90	0.009627943
BTRC	ENSP00000359206	83	136	0.009627943
CHUK	ENSP00000359424	61	92	0.009627943
GBP4	ENSP00000359490	10	11	0.009627943
GBP2	ENSP00000359497	28	39	0.009627943

GBP1	ENSP00000359504	38	47	0.009627943
IFI44	ENSP00000359783	30	34	0.009627943
IFI44L	ENSP00000359787	32	36	0.009627943
ZBP1	ENSP00000360215	11	11	0.009627943
IFIT5	ENSP00000360860	18	20	0.009627943
IFIT1	ENSP00000360869	40	54	0.009627943
IFIT3	ENSP00000360876	46	53	0.009627943
IFIT2	ENSP00000360891	36	51	0.009627943
EXOSC2	ENSP00000361433	20	32	0.009627943
SRSF3	ENSP00000362820	64	120	0.009627943
NCBP1	ENSP00000364289	108	219	0.009627943
DIS3	ENSP00000366997	16	25	0.009627943
NUP160	ENSP00000367721	75	119	0.009627943
ISG15	ENSP00000368699	51	77	0.009627943
DDX58	ENSP00000369213	71	75	0.009627943
SAMD9	ENSP00000369292	11	12	0.009627943
TRIM22	ENSP00000369299	17	18	0.009627943
RPP40	ENSP00000369391	10	11	0.009627943
IRF4	ENSP00000370343	23	57	0.009627943
GTF3A	ENSP00000370532	21	25	0.009627943
RSAD2	ENSP00000371471	32	41	0.009627943
EXOSC8	ENSP00000374354	24	40	0.009627943
ERCC2	ENSP00000375809	53	81	0.009627943
DDX60	ENSP00000377344	27	29	0.009627943
MT1X	ENSP00000377995	6	6	0.009627943
IRF9	ENSP00000380073	53	71	0.009627943
MX1	ENSP00000381599	45	57	0.009627943
IFITM1	ENSP00000386187	18	33	0.009627943
EXOSC6	ENSP00000398597	29	41	0.009627943

MAVS	ENSP00000401980	32	40	0.009627943
POLR2F	ENSP00000403852	116	196	0.009627943
ATP5D	ENSP00000215375	47	69	0.013925161
RIPK3	ENSP00000216274	9	12	0.013925161
AURKA	ENSP00000216911	111	242	0.013925161
POLR2C	ENSP00000219252	102	179	0.013925161
MEFV	ENSP00000219596	9	17	0.013925161
NOD1	ENSP00000222823	20	25	0.013925161
CHMP5	ENSP00000223500	11	24	0.013925161
NUP88	ENSP00000225696	35	59	0.013925161
GTF2H3	ENSP00000228955	40	67	0.013925161
NDUFS1	ENSP00000233190	34	54	0.013925161
TNFAIP3	ENSP00000237289	26	49	0.013925161
SRSF6	ENSP00000244020	60	112	0.013925161
BST2	ENSP00000252593	9	11	0.013925161
SNRPA1	ENSP00000254193	66	123	0.013925161
NUPL2	ENSP00000258742	35	59	0.013925161
IDH1	ENSP00000260985	11	24	0.013925161
TCEB2	ENSP00000262306	66	104	0.013925161
MCM6	ENSP00000264156	86	177	0.013925161
NDUFS6	ENSP00000274137	39	55	0.013925161
NFKBIE	ENSP00000275015	11	17	0.013925161
UPF3B	ENSP00000276201	91	188	0.013925161
NUP205	ENSP00000285968	39	62	0.013925161
PLK1	ENSP00000300093	143	318	0.013925161
POLR2G	ENSP00000301788	102	174	0.013925161
PRPF8	ENSP00000304350	65	124	0.013925161
NUP62	ENSP00000305503	43	77	0.013925161
NUP93	ENSP00000310668	39	66	0.013925161

CFLAR	ENSP00000312455	24	45	0.013925161
POLR2B	ENSP00000312735	109	190	0.013925161
MDH2	ENSP00000327070	47	67	0.013925161
TBK1	ENSP00000329967	33	46	0.013925161
UBA7	ENSP00000333266	8	8	0.013925161
GTF2F2	ENSP00000340823	63	125	0.013925161
TRADD	ENSP00000341268	32	64	0.013925161
C1D	ENSP00000348107	14	20	0.013925161
SRSF2	ENSP00000350877	68	135	0.013925161
SUPV3L1	ENSP00000352678	12	16	0.013925161
TPR	ENSP00000356448	39	67	0.013925161
EXOSC1	ENSP00000359939	18	29	0.013925161
NUP188	ENSP00000361658	35	59	0.013925161
SDHB	ENSP00000364649	36	49	0.013925161
LSM2	ENSP00000364813	14	15	0.013925161
NDUFB6	ENSP00000369176	31	49	0.013925161
SRSF5	ENSP00000377892	61	118	0.013925161
PSMD14	ENSP00000386541	60	144	0.013925161
UBA52	ENSP00000388107	171	390	0.013925161
RPP30	ENSP00000389182	9	11	0.013925161
MT1G	ENSP00000391397	5	6	0.013925161
SLBP	ENSP00000417686	38	67	0.013925161
NDUFAB1	ENSP00000007516	43	65	0.018393683
TBCB	ENSP00000221855	4	4	0.018393683
NAA38	ENSP00000249299	44	49	0.018393683
RFC1	ENSP00000261424	34	41	0.018393683
HAT1	ENSP00000264108	31	51	0.018393683
MT1E	ENSP00000307706	4	5	0.018393683
HNRNPA0	ENSP00000316042	39	89	0.018393683

SNRNP200	ENSP00000317123	61	121	0.018393683
CYC1	ENSP00000317159	49	75	0.018393683
SRSF7	ENSP00000325905	57	110	0.018393683
AMER1	ENSP00000329117	22	51	0.018393683
PSMG1	ENSP00000329915	9	18	0.018393683
NDUFB1	ENSP00000330787	28	46	0.018393683
MT1B	ENSP00000334998	4	4	0.018393683
PARK2	ENSP00000355865	43	78	0.018393683
IL6R	ENSP00000357470	16	33	0.018393683
PSMD4	ENSP00000357879	23	74	0.018393683
PRPF4	ENSP00000363313	61	120	0.018393683
CUL2	ENSP00000363880	27	42	0.018393683
UQCRQ	ENSP00000367934	50	79	0.018393683
IFNA2	ENSP00000369554	20	45	0.018393683
PIGA	ENSP00000369820	14	18	0.018393683
DCP2	ENSP00000373715	22	37	0.018393683
GTF2F1	ENSP00000377969	67	135	0.018393683
CD74	ENSP00000009530	32	46	0.021247875
PSMA3	ENSP00000216455	52	126	0.021247875
SRSF1	ENSP00000258962	69	138	0.021247875
NDUFAF1	ENSP00000260361	26	40	0.021247875
POF1B	ENSP00000262753	5	6	0.021247875
SETD1B	ENSP00000267197	12	17	0.021247875
NDUFB9	ENSP00000276689	43	66	0.021247875
TCEB1	ENSP00000284811	62	122	0.021247875
REL	ENSP00000295025	19	40	0.021247875
RPP14	ENSP00000295959	9	10	0.021247875
NDUFS4	ENSP00000296684	40	58	0.021247875
FADD	ENSP00000301838	31	64	0.021247875

POP7	ENSP00000304353	6	7	0.021247875
SF3B2	ENSP00000318861	58	112	0.021247875
NFKBIZ	ENSP00000325663	6	6	0.021247875
PNRC1	ENSP00000336931	2	2	0.021247875
CPSF1	ENSP00000339353	62	125	0.021247875
TLR9	ENSP00000353874	31	43	0.021247875
NSL1	ENSP00000355944	56	66	0.021247875
RNF2	ENSP00000356480	29	60	0.021247875
SDHC	ENSP00000356953	26	34	0.021247875
ATP5F1	ENSP00000358737	28	57	0.021247875
SRSF11	ENSP00000359988	66	130	0.021247875
RAE1	ENSP00000360286	64	104	0.021247875
CTPS1	ENSP00000361699	37	53	0.021247875
SRSF4	ENSP00000362900	60	117	0.021247875
DNAJC3	ENSP00000365991	10	17	0.021247875
EXOSC10	ENSP00000366135	30	49	0.021247875
NFX1	ENSP00000368856	56	108	0.021247875
MAP3K1	ENSP00000382423	43	66	0.021247875
RELA	ENSP00000384273	89	236	0.021247875
BIRC2	ENSP00000227758	41	52	0.024946898
NDUFA2	ENSP00000252102	37	54	0.024946898
TAF4	ENSP00000252996	41	75	0.024946898
COQ9	ENSP00000262507	10	10	0.024946898
PSMB6	ENSP00000270586	63	102	0.024946898
RFC4	ENSP00000296273	110	239	0.024946898
CEBPB	ENSP00000305422	70	117	0.024946898
IL8	ENSP00000306512	81	137	0.024946898
PAAF1	ENSP00000311665	12	21	0.024946898
RNF135	ENSP00000328340	10	13	0.024946898

LIG4	ENSP00000349393	22	25	0.024946898
EXOSC9	ENSP00000368984	21	31	0.024946898
IFNB1	ENSP00000369581	34	49	0.024946898
TNF	ENSP00000398698	133	239	0.024946898
SMG1	ENSP00000402515	13	19	0.024946898
NDUFB4	ENSP00000184266	36	53	0.027797972
UQCRC1	ENSP00000203407	37	67	0.027797972
KMT2B	ENSP00000222270	13	20	0.027797972
SRSF9	ENSP00000229390	70	114	0.027797972
HSPB1	ENSP00000248553	25	66	0.027797972
MNAT1	ENSP00000261245	46	87	0.027797972
AKT3	ENSP00000263826	49	57	0.027797972
NDUFB10	ENSP00000268668	36	53	0.027797972
TP53	ENSP00000269305	239	583	0.027797972
NXF1	ENSP00000294172	40	73	0.027797972
UQCRFS1	ENSP00000306397	36	65	0.027797972
BID	ENSP00000318822	17	29	0.027797972
ALYREF	ENSP00000331817	64	128	0.027797972
PARK7	ENSP00000340278	27	44	0.027797972
SF3A3	ENSP00000362110	53	102	0.027797972
APEX2	ENSP00000364126	11	13	0.027797972
PCNA	ENSP00000368438	136	311	0.027797972
IFNA8	ENSP00000369553	13	20	0.027797972
POMP	ENSP00000370222	14	35	0.027797972
CCNH	ENSP00000256897	50	96	0.030809419
PSMA6	ENSP00000261479	62	101	0.030809419
MCM2	ENSP00000265056	84	180	0.030809419
NDUFA9	ENSP00000266544	40	60	0.030809419
RB1	ENSP00000267163	86	181	0.030809419

POLR2D	ENSP00000272645	98	169	0.030809419
RPP25	ENSP00000317691	7	8	0.030809419
PHF6	ENSP00000329097	17	25	0.030809419
NDUFA12	ENSP00000330737	31	54	0.030809419
MYLIP	ENSP00000349298	7	8	0.030809419
GTF3C5	ENSP00000361180	20	25	0.030809419
IDH3B	ENSP00000370223	12	29	0.030809419
SETD2	ENSP00000386759	16	34	0.030809419
NEK1	ENSP00000408020	7	7	0.030809419
GTPBP1	ENSP00000216044	11	15	0.03330748
RNF125	ENSP00000217740	5	7	0.03330748
NDUFA10	ENSP00000252711	30	52	0.03330748
AMMECR1	ENSP00000262844	25	29	0.03330748
GEMIN6	ENSP00000281950	31	44	0.03330748
CASP2	ENSP00000312664	22	34	0.03330748
TAF7	ENSP00000312709	34	58	0.03330748
PKM	ENSP00000320171	78	135	0.03330748
DIS3L	ENSP00000321711	11	16	0.03330748
CUL1	ENSP00000326804	55	140	0.03330748
ATL1	ENSP00000351155	5	7	0.03330748
NDUFS2	ENSP00000356972	30	52	0.03330748
MAP1LC3A	ENSP00000363970	11	15	0.03330748
TPP2	ENSP00000365233	18	41	0.03330748
POLA1	ENSP00000368349	90	158	0.03330748
CASP4	ENSP00000388566	13	14	0.03330748
EIF2AK1	ENSP00000199389	21	24	0.03660525
TNFSF10	ENSP00000241261	24	32	0.03660525
PSMB1	ENSP00000262193	42	99	0.03660525
RNPS1	ENSP00000315859	66	170	0.03660525

EBAG9	ENSP00000337675	3	9	0.03660525
ATF3	ENSP00000344352	24	49	0.03660525
NDUFA8	ENSP00000362873	38	69	0.03660525
PTPN18	ENSP00000175756	7	8	0.039373059
CMPK2	ENSP00000256722	17	39	0.039373059
TOPBP1	ENSP00000260810	53	72	0.039373059
AP5Z1	ENSP00000297562	11	20	0.039373059
UQCRH	ENSP00000309565	45	71	0.039373059
RNASEH1	ENSP00000313350	8	10	0.039373059
SRRM1	ENSP00000326261	58	114	0.039373059
USP16	ENSP00000334808	14	22	0.039373059
CDC16	ENSP00000348554	63	128	0.039373059
POP5	ENSP00000350098	8	10	0.039373059
TOMM22	ENSP00000216034	10	23	0.041839952
SKP1	ENSP00000231487	83	146	0.041839952
AP5S1	ENSP00000246041	11	16	0.041839952
LSM7	ENSP00000252622	30	42	0.041839952
AP5M1	ENSP00000261558	11	16	0.041839952
DHX38	ENSP00000268482	62	126	0.041839952
TAP1	ENSP00000346206	13	21	0.041839952
ING1	ENSP00000364929	12	24	0.041839952
XRN2	ENSP00000366396	12	24	0.041839952
TOMM5	ENSP00000384411	10	23	0.041839952
LSM5	ENSP00000410758	17	32	0.041839952
IDH3G	ENSP00000217901	14	30	0.044013456
TBP	ENSP00000230354	111	142	0.044013456
CPSF3	ENSP00000238112	73	124	0.044013456
MAD2L1	ENSP00000296509	105	235	0.044013456
NDUFA7	ENSP00000301457	34	50	0.044013456

EIF4G1	ENSP00000316879	66	172	0.044013456
KAT5	ENSP00000340330	36	105	0.044013456
USP1	ENSP00000343526	37	55	0.044013456
BRCA1	ENSP00000350283	126	285	0.044013456
RPA2	ENSP00000363021	36	66	0.044013456
E2F4	ENSP00000368686	37	55	0.044013456
TIRAP	ENSP00000376445	16	35	0.044013456
PYCARD	ENSP00000247470	9	19	0.046572378
SAP130	ENSP00000259235	19	29	0.046572378
ANAPC5	ENSP00000261819	45	87	0.046572378
KAT2B	ENSP00000263754	80	173	0.046572378
PCBP1	ENSP00000305556	50	98	0.046572378
HNRNPD	ENSP00000313199	46	112	0.046572378
SMC1A	ENSP00000323421	68	142	0.046572378
DCLRE1C	ENSP00000367527	15	18	0.046572378
ZW10	ENSP00000200135	77	96	0.048876132
MSH6	ENSP00000234420	49	66	0.048876132
PSMA5	ENSP00000271308	60	100	0.048876132
CENPC	ENSP00000273853	58	70	0.048876132
POP1	ENSP00000339529	5	5	0.048876132
DHX9	ENSP00000356520	56	113	0.048876132
GTF3C4	ENSP00000361219	19	24	0.048876132
DNAJC9	ENSP00000362041	6	8	0.048876132
TXN	ENSP00000363641	37	56	0.048876132

Appendix B

LEAN result: Enrichment analysis

Table B.1: GSEA Enrichment results for LEAN star centers with "Hallmark" dataset as background. K is the number of genes in gene set, and k is the number of genes in overlap. The significance is measured by the FDR q-value.

Gene Set Name	Description	k	K	FDR q-value
HALLMARK INTERFERON GAMMA RESPONSE	Genes up-regulated in response to IFNG [GeneID=3458].	51	200	2.09E-61
HALLMARK INTERFERON ALPHA RESPONSE	Genes up-regulated in response to alpha interferon proteins.	37	97	4.29E-52
HALLMARK OXIDATIVE PHOSPHORYLATION	Genes encoding proteins involved in oxidative phosphorylation.	36	200	1.65E-37
HALLMARK MYC TARGETS V1	A subgroup of genes regulated by MYC - version 1 (v1).	31	200	2.89E-30
HALLMARK DNA REPAIR	Genes involved in DNA repair.	27	150	2.7E-28
HALLMARK E2F TARGETS	Genes encoding cell cycle related targets of E2F transcription factors.	24	200	6.96E-21
HALLMARK TNFA SIGNALING VIA NFKB	Genes regulated by NF-kB in response to TNF [GeneID=7124].	21	200	3.64E-17
HALLMARK APOPTOSIS	Genes mediating programmed cell death (apoptosis) by activation of caspases.	18	161	2.55E-15
HALLMARK G2M CHECKPOINT	Genes involved in the G2/M checkpoint as in progression through the cell division cycle.	16	200	1.86E-11
HALLMARK ADIPOGENESIS	Genes up-regulated during adipocyte differentiation (adipogenesis).	14	200	2.04E-9
HALLMARK INFLAMMATORY RESPONSE	Genes defining inflammatory response.	14	200	2.04E-9

HALLMARK UNFOLDED PROTEIN RESPONSE	Genes up-regulated during unfolded protein response a cellular stress response related to the endoplasmic reticulum.	11	113	4.33E-9
HALLMARK ALLOGRAFT REJECTION	Genes up-regulated during transplant rejection.	12	200	1.59E-7
HALLMARK MTORC1 SIGNALING	Genes up-regulated through activation of mTORC1 complex.	12	200	1.59E-7
HALLMARK UV RESPONSE UP	Genes up-regulated in response to ultraviolet (UV) radiation.	8	158	9.21E-5
HALLMARK PI3K AKT MTOR SIGNALING	Genes up-regulated by activation of the PI3K/AKT/mTOR pathway.	6	105	4.55E-4
HALLMARK REACTIVE OXIGEN SPECIES PATHWAY	Genes up-regulated by reactive oxigen species (ROS).	4	49	1.42E-3
HALLMARK IL6 JAK STAT3 SIGNALING	Genes up-regulated by IL6 [GeneID=3569] via STAT3 [GeneID=6774] e.g. during acute phase response.	5	87	1.42E-3
HALLMARK HYPOXIA	Genes up-regulated in response to low oxygen levels (hypoxia).	7	200	1.97E-3
HALLMARK IL2 STAT5 SIGNALING	Genes up-regulated by STAT5 in response to IL2 stimulation.	7	200	1.97E-3
HALLMARK P53 PATHWAY	Genes involved in p53 pathways and networks.	7	200	1.97E-3
HALLMARK FATTY ACID METABOLISM	Genes encoding proteins involved in metabolism of fatty acids.	6	158	2.9E-3
HALLMARK NOTCH SIGNALING	Genes up-regulated by activation of Notch signaling.	3	32	3.82E-3
HALLMARK COMPLEMENT	Genes encoding components of the complement system which is part of the innate immune system.	6	200	7.94E-3
HALLMARK GLYCOLYSIS	Genes encoding proteins involved in glycolysis and gluconeogenesis.	6	200	7.94E-3
HALLMARK KRAS SIGNALING DN	Genes down-regulated by KRAS activation.	6	200	7.94E-3

Table B.2: GSEA Enrichment results for LEAN star centers with the immunological signatures dataset as background. K is the number of genes in gene set, and k is the number of genes in overlap. The significance is measured by the FDR q-value.

Gene Set Name	Description	k	K	FDR q-value
GSE13484 UNSTIM VS YF17D VACCINE STIM PBMC DN	Genes down-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) versus PBMC stimulated with YF17D vaccine.	48	200	2.71E-54
GSE42724 NAIVE BCELL VS PLASMABLAST UP	Genes up-regulated in B lymphocytes: naive versus plasmablasts.	46	199	2.02E-51
GSE13485 CTRL VS DAY7 YF17D VACCINE PBMC DN	Genes down-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) versus PBMC 7 days after stimulation with YF17D vaccine.	46	200	2.02E-51
GSE13485 DAY1 VS DAY7 YF17D VACCINE PBMC DN	Genes down-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) 1 day after stimulation with YF17D vaccine versus PBMC 7 days after the stimulation.	45	200	5.47E-50
GSE13485 PRE VS POST YF17D VACCINATION PBMC DN	Genes down-regulated in comparison of peripheral blood mononuclear cells (PBMC) before vs after YF17D vaccination.	45	200	5.47E-50

GSE13485 DAY3 VS DAY7 YF17D VACCINE PBMC DN	Genes down-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) 3 days after stimulation with YF17D vaccine versus PBMC 7 days after the stimulation.	44	199	1.5E-48
GSE21927 SPLEEN C57BL6 VS 4T1 TUMOR BALBC MONOCYTES DN	Genes down-regulated in CD11b+ cells from spleen of healthy C57BL6 mice versus CD11b+ cells from tumor infiltrating monocytes of BALB/c mice bearing 4T1 mammary carcinoma.	44	200	1.5E-48
GSE37533 PPARG1 FOXP3 VS FOXP3 TRANSDUCED CD4 TCELL DN	Genes down-regulated in CD4 [GeneID=920] over-expressing: FOXP3 [GeneID=50943] and PPARg1 form of PPARG [GeneID=5468] versus FOXP3 [GeneID=50943].	44	200	1.5E-48
GSE18791 CTRL VS NEWCASTLE VIRUS DC 8H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 8 h.	43	200	4.61E-47
GSE18791 UNSTIM VS NEWCATSLE VIRUS DC 10H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 10 h versus cDCs infected with Newcastle disease virus (NDV) at 10 h.	43	200	4.61E-47
GSE18791 UNSTIM VS NEWCATSLE VIRUS DC 6H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 6 h versus cDCs infected with Newcastle disease virus (NDV) at 6 h.	43	200	4.61E-47
GSE10325 CD4 TCELL VS LUPUS CD4 TCELL DN	Genes down-regulated in comparison of healthy CD4 [GeneID=920] T cells versus systemic lupus erythematosus CD4 [GeneID=920] T cells.	42	200	1.3E-45
GSE19888 ADENOSINE A3R INH PRETREAT AND ACT BY A3R VS TCELL MEMBRANES ACT MAST CELL UP	Genes up-regulated in HMC-1 (mast leukemia) cells: incubated with the peptide ALL1 and then treated with Cl-IB-MECA [PubChem=3035850] versus stimulation by T cell membranes.	42	200	1.3E-45
GSE21360 NAIVE VS QUATERNARY MEMORY CD8 TCELL DN	Genes down-regulated inCD8 T cells: naive versus 4' memory.	42	200	1.3E-45
GSE22886 CTRL VS LPS 24H DC DN	Genes down-regulated in comparison of unstimulated dendritic cells (DC) versus 1 day DC stimulated with LPS (TLR4 agonist).	42	200	1.3E-45
GSE42021 TREG VS TCONV PLN UP	Genes up-regulated in cells from peripheral lymph nodes: T reg versus T conv.	42	200	1.3E-45
GSE21546 WT VS SAP1A KO DP THYMOCYTES UP	Genes up-regulated in untreated double positive thymocytes: wildtype versus ELK4 [GeneID=2005] knockout.	41	199	3.88E-44
GSE14000 UNSTIM VS 4H LPS DC DN	Genes down-regulated in comparison of dendritic cells (DC) before and 4 h after LPS (TLR4 agonist) stimulation.	41	200	4.58E-44

GSE18791 CTRL VS NEWCASTLE VIRUS DC 6H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 6 h.	40	200	1.51E-42
GSE42021 CD24HI VS CD24INT TREG THYMUS DN	Genes down-regulated in thymic T reg: CD24 high [GeneID=100133941] versus CD24 int [GeneID=100133941].	40	200	1.51E-42
GSE42021 TREG PLN VS TREG PRECURSORS THYMUS DN	Genes down-regulated in T reg from: peripheral lymph nodes versus thymic precursors.	40	200	1.51E-42
GSE13485 CTRL VS DAY3 YF17D VACCINE PBMC DN	Genes down-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) versus PBMC 3 days after stimulation with YF17D vaccine.	39	200	4.88E-41
GSE21360 NAIVE VS QUATERNARY MEMORY CD8 TCELL UP	Genes up-regulated inCD8 T cells: naive versus 4' memory.	39	200	4.88E-41
GSE37533 PPARG2 FOXP3 VS FOXP3 TRANSDUCED CD4 TCELL DN	Genes down-regulated in CD4 [GeneID=920] over-expressing: FOXP3 [GeneID=50943] and PPARg2 form of PPARG [GeneID=5468] versus FOXP3 [GeneID=50943].	39	200	4.88E-41
GSE14000 UNSTIM VS 4H LPS DC TRANSLATED RNA DN	Genes down-regulated in comparison of polysome bound (translated) mRNA before and 4 h after LPS (TLR4 agonist) stimulation.	38	200	1.61E-39
GSE19888 ADENOSINE A3R INH VS ACT WITH INHIBITOR PRETREATMENT IN MAST CELL UP	Genes up-regulated in HMC-1 (mast leukemia) cells incubated the peptide ALL1 versus those followed by treatment with Cl-IB-MECA [PubChem=3035850].	38	200	1.61E-39
GSE10325 MYELOID VS LUPUS MYELOID DN	Genes down-regulated in comparison of healthy myeloid cells versus systemic lupus erythematosus myeloid cells.	37	200	1.61E-39
GSE37533 PPARG1 FOXP3 VS FOXP3 TRANSDUCED CD4 TCELL PIOGLITAZONE TREATED UP	Genes up-regulated in CD4 [GeneID=920] T cells treated with pioglitazone [PubChem=4829] and over-expressing: FOXP3 [GeneID=50943] and PPARg1 isoform of PPARG [GeneID=5468] versus FOXP3 [GeneID=50943].	37	200	1.51E-38
GSE42021 CD24INT VS CD24LOW TREG THYMUS DN	Genes down-regulated in thymic T reg: CD24 int [GeneID=100133941] versus CD24 low [GeneID=100133941].	37	200	7.36E-38
GSE18791 CTRL VS NEWCASTLE VIRUS DC 4H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 4 h.	36	200	2.44E-36
GSE6269 FLU VS STREP PNEUMO INF PBMC UP	Genes up-regulated in comparison of peripheral blood mononuclear cells (PBMC) from patients with acute influenza infection versus PBMC from patients with acute S. pneumoniae infection.	34	169	3.44E-36

GSE21546 UNSTIM VS ANTI CD3 STIM ELK1 KO DP THYMOCYTES UP	Genes up-regulated in double positive thymocytes with ELK1 [GeneID=2002] knockout: untreated versus stimulated by anti-CD3.	35	196	2.29E-35
GSE24634 IL4 VS CTRL TREATED NAIVE CD4 TCELL DAY5 DN	Genes down-regulated in comparison of CD25- T cells treated with IL4 [GeneID=3565] at day 5 versus untreated CD25- T cells at day 5.	35	200	4.19E-35
GSE33424 CD161 INT VS NEG CD8 TCELL UP	Genes up-regulated in CD8 T cells: KLRB1 int [GeneID=3820] versus KLRB1- [GeneID=3820].	35	200	4.19E-35
GSE37533 PPARG1 FOXP3 VS PPARG2 FOXP3 TRANSDUCED CD4 TCELL PIOGLITAZONE TREATED DN	Genes down-regulated in CD4 [GeneID=920] T cells treated with pioglitazone [PubChem=4829] and over-expressing: FOXP3 [GeneID=50943] and PPARg1 isoform of PPARG [GeneID=5468] versus FOXP3 [GeneID=50943] and PPARg2 form of PPARG [GeneID=5468].	35	200	4.19E-35
GSE37534 UNTREATED VS PIOGLITAZONE TREATED CD4 TCELL PPARG1 AND FOXP3 TRASDUCED DN	Genes down-regulated in CD4 [GeneID=920] T cells over-expressing FOXP3 [GeneID=50943] and PPARg1 form of PPARG [GeneID=5468]: untreated versus pioglitazone [PubChem=4829].	35	200	4.19E-35
GSE42021 CD24HI VS CD24LOW TREG THYMUS DN	Genes down-regulated in thymic T reg: CD24 high [GeneID=100133941] versus CD24 low [GeneID=100133941].	35	200	4.19E-35
GSE19888 ADENOSINE A3R INH VS TCELL MEMBRANES ACT MAST CELL UP	Genes up-regulated in HMC-1 (mast leukemia) cells: incubated with the peptide ALL1 versus stimulated with T cell membranes.	34	199	1.06E-33
GSE18791 CTRL VS NEWCASTLE VIRUS DC 10H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 10 h.	34	200	1.14E-33
GSE34205 HEALTHY VS FLU INF INFANT PBMC DN	Genes down-regulated in comparison of peripheral blood mononuclear cells (PBMC) from healthy donors versus PBMCs from infanct with acute influenza infection.	34	200	1.14E-33
GSE36527 CD69 NEG VS POS TREG CD62L LOS KLRG1 NEG UP	Genes up-regulated in KLRG1- SELL low [GeneID=10219 and 6402] T reg: CD69- [GeneID=969] versus CD69+ [GeneID=969].	34	200	1.14E-33
GSE42021 TREG PLN VS CD24INT TREG THYMUS DN	Genes down-regulated in T reg: peripheral lymph nodes versus thymic CD24 int [GeneID=100133941].	34	200	1.14E-33
GSE2770 TGFB AND IL4 ACT VS ACT CD4 TCELL 2H DN	Genes down-regulated in CD4 [GeneID=920] T cells activated by anti-CD3 and anti-CD28: TGFB1 and IL4 [GeneID=7040 and 3565] (2h) versus untreated (2h).	33	199	2.8E-32
GSE13485 DAY7 VS DAY21 YF17D VACCINE PBMC UP	Genes up-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) 7 days after stimulation with YF17D vaccine versus PBMC 21 days after the stimulation.	33	200	3.04E-32

GSE18281 CORTICAL VS MEDULLARY THYMOCYTE UP	Genes up-regulated in thymocytes: cortical versus medullary sources.	33	200	3.04E-32
GSE26030 TH1 VS TH17 DAY5 POST POLARIZATION UP	Genes up-regulated in T helper cells 5 days post polarization: Th1 versus Th17.	33	200	3.04E-32
GSE34006 A2AR KO VS A2AR AGONIST TREATED TREG UP	Genes up-regulated in T reg: untreated ADORA2A [GeneID=135] knockout versus wildtype treated by ZM 241385 [PubChem=176407].	33	200	3.04E-32
GSE34156 UNTREATED VS 6H NOD2 AND TLR1 TLR2 LIGAND TREATED MONOCYTE DN	Genes down-regulated in monocytes (6h): untreated versus muramyl dipeptide [PubChem=11620162] andM. tuberculosis 19 kDa lipopeptide.	30	154	1.59E-31
GSE40685 TREG VS FOXP3 KO TREG PRECURSOR UP	Genes up-regulated in CD4: FOXP3+ [GeneID=50943] T reg versus FOXP3 [GeneID=50943] knockout T reg precursor.	32	195	3.58E-31
GSE10325 BCELL VS LUPUS BCELL DN	Genes down-regulated in comparison of healthy B cells versus systemic lupus erythematosus B cells.	32	200	7.59E-31
GSE1432 CTRL VS IFNG 24H MICROGLIA DN	Genes down-regulated in comparison of control microglia cells versus those 24 h after stimulation with IFNG [GeneID=3458].	32	200	7.59E-31
GSE22140 GERMFREE VS SPF MOUSE CD4 TCELL UP	Genes up-regulated in healthy CD4 [GeneID=920] T cells: germ free versus specific pathogen free.	32	200	7.59E-31
GSE26890 CXCR1 NEG VS POS EFFECTOR CD8 TCELL UP	Genes up-regulated in effector CD8 T cells: CXCR1+ [GeneID=3577] versus CXCR1- [GeneID=3577].	32	200	7.59E-31
GSE42021 TREG PLN VS CD24LO TREG THYMUS DN	Genes down-regulated in T reg: peripheral lymph nodes versus thymic CD24 low [GeneID=100133941].	32	200	7.59E-31
GSE18791 UNSTIM VS NEWCATSLE VIRUS DC 18H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 18 h versus cDCs infected with Newcastle disease virus (NDV) at 18 h.	31	200	1.98E-29
GSE38681 WT VS LYL1 KO LYMPHOID PRIMED MULTIPOTENT PROGENITOR DN	Genes down-regulated in lymphoid primed multipotent progenitors: wildtype versus LYL1 [GeneID=4066] knockout.	31	200	1.98E-29
GSE42021 CD24INT VS CD24LOW TCONV THYMUS DN	Genes down-regulated in thymic T conv: CD24 int [GeneID=100133941] versus CD24 low [GeneID=100133941].	31	200	1.98E-29
GSE21546 ELK1 KO VS SAP1A KO AND ELK1 KO DP THYMOCYTES UP 4.4E-28	Genes up-regulated in untreated double positive thymocytes: ELK1 [GeneID=2002] knockout versus ELK1 and ELK4 [GeneID=2002	2005] knock-out.	30	199

GSE34006 WT VS A2AR KO TREG DN	Genes down-regulated in T reg: wildtype versus ADORA2A [GeneID=135].	30	200	4.89E-28
GSE41978 ID2 KO VS BIM KO KLRG1 LOW EFFECTOR CD8 TCELL UP	Genes up-regulated in KLRG1 low [GeneID=10219] CD8 T effector cells during infection: ID2 [GeneID=10219] knockout versus BCL2L11 [GeneID=10018] knockout.	30	200	4.89E-28
GSE7548 NAIVE VS DAY7 PCC IMMUNIZATION CD4 TCELL DN	Genes down-regulated in CD4 [GeneID=920] T cells from lymph nodes: naive versus day 7 after immunization.	30	200	4.89E-28
GSE1432 CTRL VS IFNG 6H MICROGLIA DN	Genes down-regulated in comparison of control microglia cells versus those 6 h after stimulation with IFNG [GeneID=3458].	29	200	1.22E-26
GSE18791 CTRL VS NEWCASTLE VIRUS DC 16H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 16 h.	28	200	2.82E-25
GSE2706 UNSTIM VS 2H LPS AND R848 DC DN	Genes down-regulated in comparison of unstimulated dendritic cells (DC) at 0 h versus DCs stimulated with LPS (TLR4 agonist) and R848 for 2 h.	28	200	2.82E-25
GSE2706 UNSTIM VS 8H R848 DC DN	Genes down-regulated in comparison of unstimulated dendritic cells (DC) at 0 h versus DCs stimulated with R848 for 8 h.	28	200	2.82E-25
GSE18791 CTRL VS NEWCASTLE VIRUS DC 12H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 12 h.	27	200	6.47E-24
GSE6269 FLU VS E COLI INF PBMC UP	Genes up-regulated in comparison of peripheral blood mononuclear cells (PBMC) from patients with acute influenza infection versus PBMC from patients with acute E. coli infection.	25	162	1.35E-23
GSE6269 HEALTHY VS STAPH AUREUS INF PBMC UP	Genes up-regulated in comparison of peripheral blood mononuclear cells (PBMC) from patients with acute influenza infection versus PBMC from patients with acute S. aureus infection.	25	171	5.41E-23
GSE14000 UNSTIM VS 16H LPS DC TRANSLATED RNA DN	Genes down-regulated in comparison of polysome bound (translated) mRNA before and 16 h after LPS (TLR4 agonist) stimulation.	26	200	1.28E-22
GSE1432 1H VS 6H IFNG MICROGLIA DN	Genes down-regulated in comparison of microglia cells 1 h after stimulation with IFNG [GeneID=3458] versus microglia cells 6 h after the stimulation.	26	200	1.28E-22
GSE18281 SUBCAPSULAR VS CENTRAL CORTICAL REGION OF THYMUS DN	Genes down-regulated in thymus cortical regions: subcapsular versus central cortical.	26	200	1.28E-22

GSE2706 R848 VS R848 AND LPS 2H STIM DC DN	Genes down-regulated in comparison of dendritic cells (DC) stimulated with R848 at 2 h versus DCs stimulated with LPS (TLR4 agonist) and R848 for 2 h.	26	200	1.28E-22
GSE37301 MULTIPOTENT PROGENITOR VS GRAN MONO PROGENITOR DN	Genes down-regulated in multipotent progenitors versus granulocyte-monocyte progenitors.	26	200	1.28E-22
GSE7509 DC VS MONOCYTE WITH FCGRIIB STIM DN	Genes down-regulated in response to anti-FcgRIIB: dendritic cells versus monocytes.	26	200	1.28E-22
GSE7218 IGM VS IGG SIGNAL THGOUGH ANTIGEN BCELL DN	Genes down-regulated in B lymphocytes treated by anti-HEL and expressing BCR [GeneID=613] fusions with: IgM versus IgMG.	24	170	1.03E-21
GSE21546 UNSTIM VS ANTI CD3 STIM SAP1A KO AND ELK1 KO DP THYMOCYTES UP	Genes up-regulated in double positive thymocytes with ELK1 and ELK4 [GeneID=2002 and 2005] knockout: untreated versus stimulated by anti-CD3.	25	196	1.59E-21
GSE17974 IL4 AND ANTI IL12 VS UNTREATED 24H ACT CD4 TCELL DN	Genes down-regulated in comparison of CD4 [GeneID=920] T cells treated with IL4 [GeneID=3565] and anti-IL12 at 24 h versus the untreated cells at 24 h.	25	200	2.58E-21
GSE42021 CD24HI VS CD24LOW TCONV THYMUS DN	Genes down-regulated in thymic T conv: CD24 high [GeneID=100133941] versus CD24 low [GeneID=100133941].	25	200	2.58E-21
GSE13485 DAY1 VS DAY3 YF17D VACCINE PBMC DN	Genes down-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) 1 day after stimulation with YF17D vaccine versus PBMC 3 days after the stimulation.	24	200	4.78E-20
GSE15930 STIM VS STIM AND TRICHOSTATINA 48H CD8 T CELL DN	Genes down-regulated in comparison of unstimulated CD8 T cells at 48 h versus CD8 T cells at 48 h after treatment with trichostatin A (TSA) [PubChem=5562].	24	200	4.78E-20
GSE22140 GERMFREE VS SPF ARTHRITIC MOUSE CD4 TCELL UP	Genes up-regulated in arthritic (KRN model) CD4 [GeneID=920] T cells: germ free versus specific pathogen free conditions.	24	200	4.78E-20
GSE36826 WT VS IL1R KO SKIN STAPH AUREUS INF DN	Genes down-regulated in lesional skin biopsies after S. aureus infection: wildtype versus IL1R1 [GeneID=3554].	24	200	4.78E-20
GSE3982 CTRL VS LPS 48H DC DN	Genes down-regulated in comparison of untreated dendritic cells (DC) versus DCs treated with LPS (TLR4 agonist) at 48 h.	24	200	4.78E-20
GSE3982 CTRL VS LPS 4H MAC DN	Genes down-regulated in comparison of untreated macrophages versus macrophages treated with LPS (TLR4 agonist) at 4 h.	24	200	4.78E-20

GSE43863 TFH VS LY6C LOW CXCR5NEG EFFECTOR CD4 TCELL UP	Genes up-regulated in CD4 [GeneID=920] SMARTA effector T cells during acute infection of LCMV: follicular helper (Tfh) versus Ly6c low CXCR5- [GeneID=643].	24	200	4.78E-20
GSE9960 HEALTHY VS GRAM POS SEPSIS PBMC DN	Genes down-regulated in peripheral blood monocytes (PMBC):healthy versus Gram positive sepsis.	23	198	7.22E-19
GSE15930 STIM VS STIM AND IFNAB 48H CD8 T CELL DN	Genes down-regulated in comparison of unstimulated CD8 T cells at 48 h versus CD8 T cells at 48 h after stimulation with antigen-B7-1.	23	200	8.32E-19
GSE16755 CTRL VS IFNA TREATED MAC DN	Genes down-regulated in comparison of control macrophages versus macrophages treated with interferon alpha.	23	200	8.32E-19
GSE19888 ADENOSINE A3R ACT VS A3R ACT WITH A3R INH PRETREATMENT IN MAST CELL DN	Genes down-regulated in HMC-1 (mast leukemia) cells: Cl-IB-MECA [PubChem=3035850] versus incubated with the ALL1 peptide followed by treatment with Cl-IB-MECA [PubChem=3035850].	23	200	8.32E-19
GSE25085 FETAL BM VS ADULT BM SP4 THYMIC IMPLANT UP	Genes up-regulated in thymic implants from fetal versus those from adult bone marrow.	23	200	8.32E-19
GSE2706 UNSTIM VS 2H LPS DC DN	Genes down-regulated in comparison of unstimulated dendritic cells (DC) at 0 h versus DCs stimulated with LPS (TLR4 agonist) for 2 h.	23	200	8.32E-19
GSE34156 NOD2 LIGAND VS TLR1 TLR2 LIGAND 6H TREATED MONOCYTE UP	Genes up-regulated in monocytes (6h): muramyl dipeptide [PubChem=11620162] versus M. tuberculosis 19 kDa lipopeptide.	23	200	8.32E-19
GSE37534 UNTREATED VS ROSIGLITAZONE TREATED CD4 TCELL PPARG1 AND FOXP3 TRASDUCED DN	Genes down-regulated in CD4 [GeneID=920] T ceels over-expressing FOXP3 [GeneID=920] and PPARg1 isoform of PPARG [GeneID=5468]: untreated versus rosiglitazone [PubChem=77999].	23	200	8.32E-19
GSE39382 IL3 VS IL3 IL33 TREATED MAST CELL DN	Genes down-regulated in bone marrow-derived mast cells treated with IL3 [GeneID=3562]: control versus IL33 [GeneID=90865].	23	200	8.32E-19
GSE1432 1H VS 24H IFNG MICROGLIA DN	Genes down-regulated in comparison of microglia cells 1 h after stimulation with IFNG [GeneID=3458] versus microglia cells 24 h after the stimulation.	22	200	1.39E-17
GSE18791 CTRL VS NEWCASTLE VIRUS DC 14H DN	Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 14 h.	22	200	1.39E-17
GSE19888 CTRL VS TCELL MEMBRANES ACT MAST CELL PRETREAT A3R INH DN	Genes down-regulated in HMC-1 (mast leukemia) cells: untreated versus incubated with the peptide ALL1 followed by stimulation with T cell membranes.	22	200	1.39E-17

GSE21360 PRIMARY VS TERTIARY MEMORY CD8 TCELL DN	Genes down-regulated in memory CD8 T cells: 1' versus 3'.	22	200	1.39E-17
GSE22886 NAIVE CD4 TCELL VS 48H ACT TH1 DN	Genes down-regulated in comparison of naive CD4 [GeneID=920] T cells versus stimulated CD4 [GeneID=920] Th1 cells at 48 h.	22	200	1.39E-17
GSE2706 2H VS 8H R848 STIM DC DN	Genes down-regulated in comparison of dendritic cells (DC) stimulated with R848 at 2 h versus DCs stimulatd with R848 for 8 h.	22	200	1.39E-17