Computational analysis on the effects of variations in T and B cells. Primary immunodeficiencies and cancer neoepitopes

Teku, Gabriel Ndipagbornchi

2017

Document Version:
Peer reviewed version (aka post-print)

Computational analysis on the effects of variations in T and B cells

Computational analysis on the effects of variations in T and B cells

Primary immunodeficiencies and cancer neoepitopes

Gabriel Ndipagbornchi Teku

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Medicine, Lund University, Sweden.
To be defended at Segerfalksalen BMC, Lund.
On Saturday September 30 at 09:00.

Faculty opponent

Professor Olli Yli-Harja

Organization LUND UNIVERSITY	Document name DOCTORAL DISSERTATION
	Date of issue September 30th, 2017
Author Gabriel Ndipagbornchi Teku	Sponsoring organization

Title and subtitle

Computational analysis on the effects of variations in T and B cells: Primary immunodeficiencies and cancer neoepitopes

Abstract

Computational approaches are essential to study the effects of inborn and somatic variations. Results from such studies contribute to better diagnosis and therapies. Primary immunodeficiencies (PIDs) are rare inborn defects of key immune response genes. Somatic variations are main drivers of most cancers. Large and diverse data on PID genes and proteins can enable systems biology studies on their dynamic effects on T and B cells. Amino acid substitutions (AASs) are somatic variations that drive cancers. However, AASs also cause cancer-associated antigens that are recognized by lymphocytes as non-self, and are called neoantigens. Detail analysis these neoantigens can be performed due to the availability of cancer data from many consortia. The purpose of this thesis was to investigate the effects of PIDs on T and B cells and to explore features of neoepitopes in cancers. The object of the first study was to detect the central T cell-specific protein network. The purpose of the second and third studies were to reconstruct the T and B cell network model and simulate the dynamic effects of PID perturbations. The aim of the fourth study was to characterize neoepitopes from pan-cancer datasets.

The immunome interactome was reconstructed, and the links weighed with gene expression correlation of integrated, time series data (Paper I). The significance of the weighted links were computed with the Global Statistical Significance (GloSS) method, and the weighted interactome network was filtered to obtain the central T cell network. Next, the T cell network model was reconstructed from literature mining and the core T cell protein interaction network (Paper II). The B cell network model was reconstructed by mining the literature for central B cell interactions (Paper III). The normalized HillCube software was used to study the dynamic effects of PID perturbations in T and B cells. Proteome-wide amino AASs on putatively derived 8-, 9-, 10-, and 11-mer neoepitopes in 30 cancer types were analyzed with the NetMHC 4.0 software (Paper IV).

The interconnectedness of the major T cell pathways are maintained in the central T cell PPI network. Empirical evidence from Gene Ontology term and essential genes enrichment analyses were in support for the central T cell network. In the T and B cell simulations for several knockout PIDs correspond to previous results. In the T cell model, simulations for TCR, PTPRC, LCK, ZAP70 and ITK indicated profound disruption in network dynamics. BCL10, CARD11, MALT1, NEMO and MAP3K14 simulations showed significant effects. In B cell, the simulations for LYN, BTK, STIM1, ORAI1, CD19, CD21 and CD81 indicated profound changes to many proteins in the network. Severe effects were observed in the BCL10, IKKB, knockout CARD11, MALT1, NEMO, IKKB and WIPF1 simulations. No major effects were observed for constitutively active PID proteins.

The most likely epitopes are those which are detected by several macromolecular histocompatibility complexes (MHCs) and of several peptide lengths. 0.17% of all variants yield more than 100 neoepitopes. Amino acid distributions indicate that variants at all positions in neoepitopes of any length are, on average, more hydrophobic compared to the wild-type.

The core T cell network approach is general and applicable to any system with adequate data. The T and B cell models enable the understanding of the dynamic effects of PID disease processes and reveals several novel proteins that may be of interest when diagnosing and treating immunological defects. The neoepitope characteristics can be employed for targeted cancer vaccine applications in personalized therapies.

Key words		
Classification system and/or index terms (if any)		
Supplementary bibliographical information		Language: English
ISSN and key title: 1652-8220		ISBN: 978-91-7619-533-8
Recipient's notes	Number of pages	Price
	Security classification	

Signature Date 2017-08-28

Computational analysis on the effects of variations in T and B cells

Primary immunodeficiencies and cancer neoepitopes

Gabriel Ndipagbornchi Teku

Coverphoto by Gabriel Ndiagbornchi Teku

Faculty of Medicine
Department of Experimental Medical Science

Lund University, Faculty of Medicine Doctoral Dissertation Series 2017:150
ISBN 978-91-7619-533-8
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University, Lund 2017

To my wife, Dibo, and kids, Naseem and Oben
for your love, care, support and immense sacrifice.

Contents

Papers included in this thesis

Paper I

Identification of core T cell network based on immunome interactome

Teku Gabriel N, Ortutay Csaba, Vihinen Mauno

BMC Systems Biology. 2014 Feb 15; 8:17.

Paper II

Simulation of the dynamics of primary immunodeficiencies in CD4+T-cells

Teku Gabriel N, Vihinen Mauno

PLoS One. 2017 Apr 27;12(4): e0176500.

Paper III

Simulation of the dynamics of primary immunodeficiencies in B cells

Teku Gabriel N, Vihinen Mauno

(Manuscript)

Paper IV

Pan-cancer analysis of neoantigens

Teku Gabriel N, Vihinen Mauno

(Manuscript)

Abstract

Background: Computational approaches are essential to study the effects of inborn and somatic variations. Results from such studies contribute to better diagnoses and therapies. Primary immunodeficiencies (PIDs) are rare inborn defects of key immune response genes. Somatic variations are the main drivers of most cancers. Large and diverse data on PID genes and proteins can enable systems biology studies of their effects on T and B cells. Amino acid substitutions (AASs) are somatic variations that drive cancers. However, AASs also cause cancer-associated antigens that are recognized by lymphocytes as non-self, and are called neoantigens. Detail analysis these neoantigens can be performed due to the availability of cancer data from many consortia.

Aims: The purpose of this thesis was to investigate the effects of PIDs on T and B cells and to explore features of neoepitopes in cancers. The object of the first study was to detect the central T cell-specific protein network. The purpose of the second and third studies were to reconstruct the T and B cell network models and simulate the dynamic effects of PID perturbations. The aim of the fourth study was to characterize neoepitopes from pan-cancer datasets.

Methods: The immunome interactome was reconstructed, and the links weighed with gene expression correlation of integrated, time series data. The significance of the weighted links was computed with the Global Statistical Significance (GloSS) method, and the weighted interactome network was filtered to obtain the central T cell network.

The T cell network model was reconstructed from literature mining and the core T cell protein interaction network. The B cell network model was reconstructed by mining the literature for central B cell interactions. The normalized HillCube software was then used to study the dynamic effects of PID perturbations T and B cells.

Proteome-wide AASs on putatively derived 8-, 9-, 10-, and 11-mer neoepitopes in 30 cancer types were analyzed with the NetMHC 4.0 software.

Results: The interconnectedness of the major T cell pathways were maintained in the central T cell protein-protein interaction (PPI) network. Empirical evidence from Gene Ontology term and essential genes enrichment analyses were in support for the central T cell network.

In the T and B cell simulations, the results for several knockout PIDs correspond to previous results. In the T cell model, simulations for TCR PTPRC, LCK, ZAP70 and ITK indicated profound disruption in network dynamics. BCL10, CARD11, MALT1, NEMO, IKKB and MAP3K14 simulations showed significant effects.

In B cell, the simulations for LYN, BTK, STIM1, ORAI1, CD19, CD21 and CD81 indicated profound changes to many proteins in the network. Severe effects were observed in the BCL10, IKKB, knockout CARD11, MALT1, NEMO and WIPF1 simulations. No major effects were observed for constitutively active PID proteins.

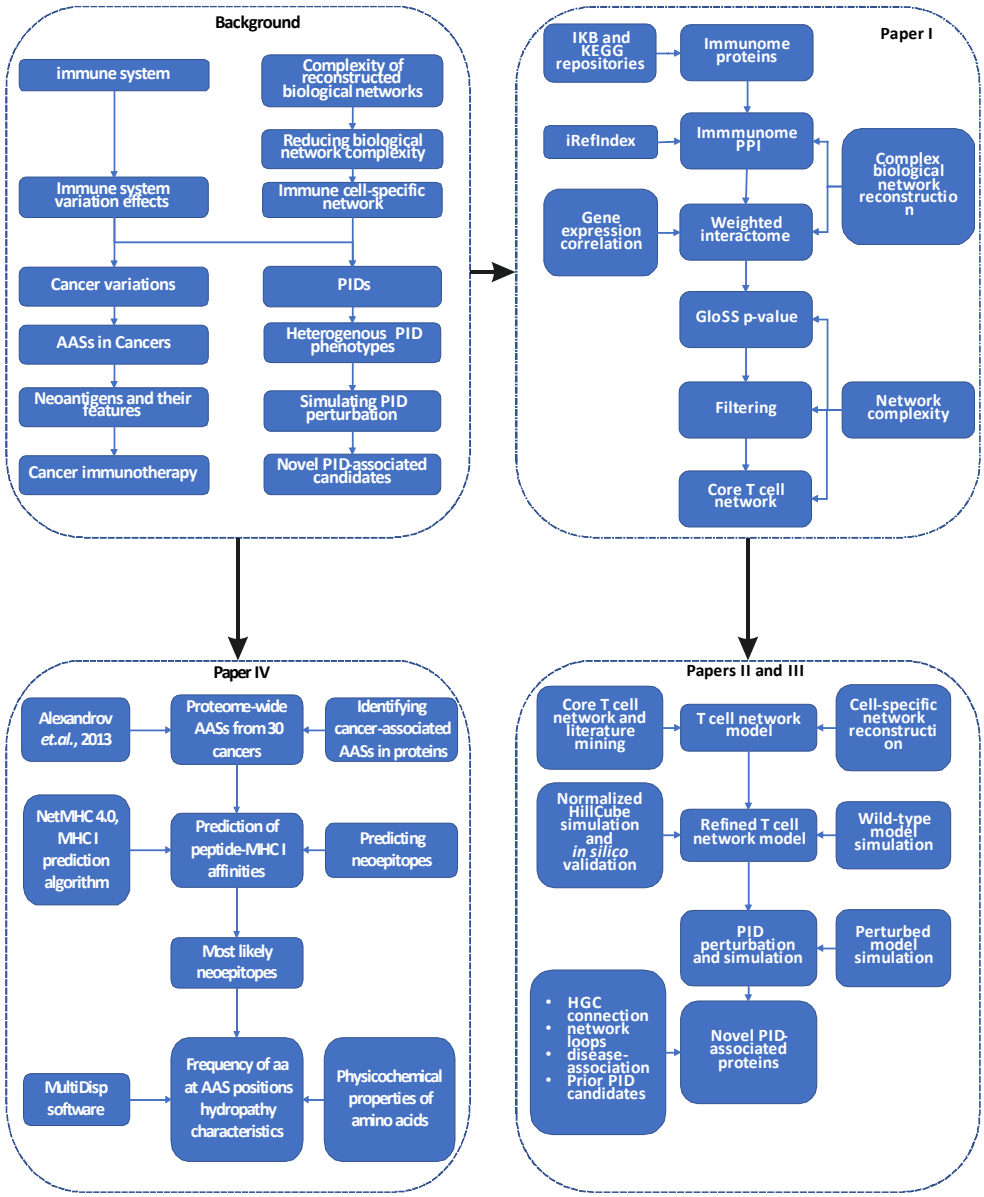
The most likely epitopes are those which are detected by several major histocompatibility complexes (MHCs) and of several peptide lengths. 0.17% of all variants yield more than 100 neoepitopes. Amino acid distributions indicate that variants at all positions in neoepitopes of any length are on average more hydrophobic compared to the wild-type.

Conclusions: The core T cell network approach is general and applicable to any system with adequate data. The T and B cell models enable the understanding of the dynamic effects of PID disease processes and reveals several novel proteins that may be of interest when diagnosing and treating immunological defects. The neoepitope characteristics can be employed for targeted cancer vaccine applications in personalized therapies.

Abbreviations

AAS	Amino acid substitution
BCR	B cell receptor
CBM complex	CARD11-BCL10-MALT1 complex
FBL	Feedback loop
FFL	Feedforward loop
GEO	Gene Expression Omnibus
GloSS	Global Statistical Significance
GO	Gene Ontology
HLA	Human leukocyte antigen
IKB	Immunome knowledge base
IUIS	International Union of Immunological Societies
KEGG	Kyoto encyclopedia of genes and genomes
MCH	Major histocompatibility complex
ODE	Ordinary differential equation
PID	Primary immunodeficiency
PPI	Protein-protein interaction
TCR	T cell receptor
Th	Helper T cells
TPPIN	T cell protein-protein interaction network

Thesis at a glance



General introduction

Networks

A network is a set of nodes connected to each other by links (Figure 1). The nodes are also called vertices (vertex for singular), and the links are also known as edges or arcs (Newman, 2010). In Figure 1, the nodes are labeled with digits. Networks are used to represent complex relations and processes between entities (Jasny, Zahn, & Marshall, 2009).

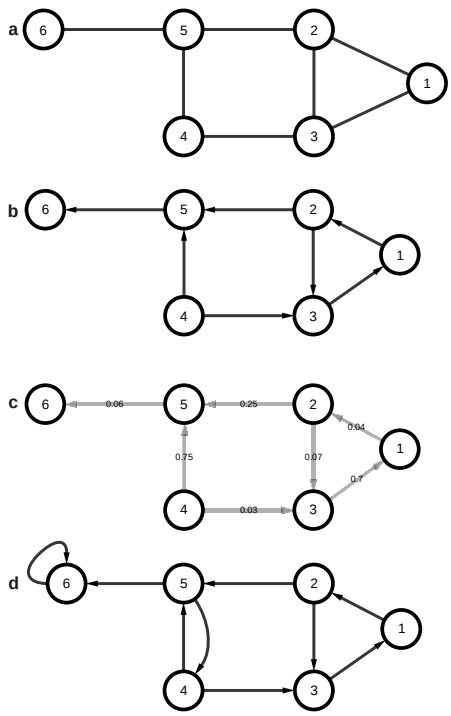


Figure 1 Different types of networks. a) undirected network with 6 nodes and 7 edges. b) directed network. c) weighted directed network. d) directed network with multiple edges.

Since many real-life problems can easily be represented as networks, this representation is usually used to model and study different aspects of complex systems (Newman, 2010). With the network representation, the nature and structure of the interactions between entities of the system can be studied using complex network theory.

To represent a problem as a network, its entities and their relationships or interactions should have a natural correspondence to the network elements, that is, nodes and links (Aldous & Wilson, 2000). Most complex systems are easily represented and modeled as networks. Examples of systems represented as complex networks include friendship, scientific collaboration, transport, the internet, genetic interaction and protein-protein interaction networks (Figure 2).

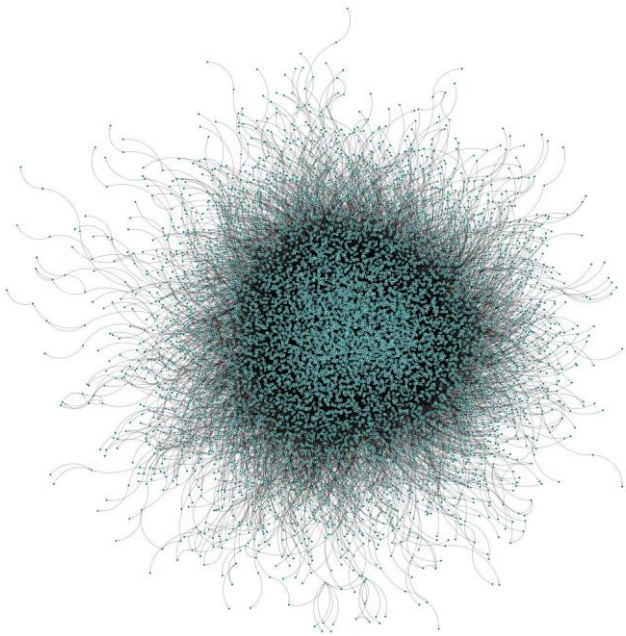


Figure 2. The immunome interactome as a hair-ball network. Due to its complexity, it is close to impossible to study such networks intuitively.

Leonhard Euler laid the foundation for the study of networks as a field in mathematics in 1736 when he solved the famous 'Seven Bridges of Königsberg' problem (Shields, 2012). Thereafter, many other scholars§ working on diverse mathematical problems furthered studies in the field, including Thomas Kirkman and William Hamilton who contributed to the existence of cycles in polyhedrons, Gustav Kirchhoff's studies on components of electrical circuits as network

elements, and works on the enumeration of chemical isomers by Arthur Cayley, James Sylvester and George Polya (Aldous & Wilson, 2000).

The structure of a network may reveal interesting and important properties of the process or system it models. The network structure is described with network measures. Network measures can be based on the nodes, links or both (local network measure). The measures can also depend on the entire network (global network measure).

There is a myriad of measures that are used to describe the structure or topology of a network. The most commonly used network measures include degree, path, cycle, connectivity, clustering coefficient and centrality (Aldous & Wilson, 2000; Newman, 2010).

The degree of a node is the number of edges connected to it. In Figure 1a, the degree of the node labeled 2, is 3. A path is a set of edges that connect a sequence of distinct vertices. In Figure 1a, a path between nodes 1 and 6 is *{1,2,5,6}*. Networks can either be undirected or directed depending on whether or not the edges have an arrow. The edges of undirected networks have no arrows (Figure 1a), whereas those of directed networks have arrows (Figure 1b). The arrows of directed networks usually depict the direction of the interactions. A network (or its part) is connected (for undirected networks) or strongly connected (for directed networks) if there is a path between every pair of nodes.

The shortest path length between a pair of nodes is the path that has the lowest number of nodes. For instance, the shortest path length between nodes 1 and 4 of Figure 1a is 2, i.e. {(1,3), (3,4)}. The diameter of a network is the longest shortest path length. A cycle is a set of connected nodes such that the start and end nodes are the same. In Figure 1b, {(1,2), (2,3), (3,1)} represent a cycle.

The clustering coefficient of a network can be defined globally or locally. The global definition of the clustering coefficient of a network indicates the degree of clustering of the nodes based on the density of node triplets (Luce & Perry, 1949). On the other hand, the local clustering coefficient of a node indicates how connected its neighbors are (Watts & Strogatz, 1998).

The centrality of a node is a measure of how important it is in the network (Newman, 2010). There are many measures of centrality, each describing a different concept of importance. Some examples include betweenness, closeness and eigenvalue centralities.

Complex networks

There are many types of network models, ranging from regular to random networks. Regular networks are those for which the degree of each node is the same (Aldous & Wilson, 2000). Examples of regular networks include those whose nodes and links form a triangle, a rectangle, a pentagon, and many other polygons. In Figure 1b, the subnetwork formed by the set of nodes *{1,2,3}* and *{2,3,4,5}*, are regular.

Random networks are constructed by starting with isolated nodes and then adding edges using a probability function (Bollobás, 2001). Many random network models exist, differing from others by the probability function used in connecting the nodes. Among the most common is the Erdős–Rényi model (Bollobás, 2001; Erdos & Renyi, 1960). Regular and random networks are similar in that the degree distribution (number of nodes with degree k, where $k = 0, 1, 2, 3\ldots$) of the nodes of the network is similar to the average degree of the network.

In complex networks, the properties are nontrivial and differ significantly from those of regular and random networks (Albert & Barabasi, 2002). For example, the degree distribution of the nodes does not follow any scale. The two types of complex networks are the scale-free and the small-world models.

In scale-free networks, the degree distribution of the nodes has a heavy tail. In other words, most of the nodes have a low degree (Caldarelli, 2007). The node distribution and clustering coefficient follows a power-law, $P(k) = k^{-\gamma}$, where $2<\gamma<3$. The nodes with the largest degree are called hubs, which are very important for the robustness and fault tolerance of the network. The average distance between the nodes of scale-free networks is very small compared to those of regular and random networks.

Small-world networks are characterized by short path lengths between nodes, high clustering coefficients and small network diameters (Watts & Strogatz, 1998). A wide variety of networks, including some random and empirical networks (e.g. metabolic networks), have these small-world properties.

Biological networks

Biological networks are complex networks that apply to biological systems and processes (de Silva & Stumpf, 2005). They include ecological (Ings et al., 2009), evolutionary (Braun et al., 2011), physiological and neural networks (Hopfield, 1982; Pal, Papp, & Lercher, 2005). The modeling of diseases as complex networks has led to the fast-growing field of network medicine (Barabasi, Gulbahce, & Loscalzo, 2011; Goh et al., 2007).

There has been a shift in recent years from focused studies of single genes and proteins to large-scale studies enabled by complex network science methods. These are denoted by the "omics" or "ome" (Joyce & Palsson, 2006). These include genome, proteome, phenome, diseasome and interactome. These networks have similar properties to other complex networks.

There is a wide variety of biological networks. These include protein-protein interaction (PPI), gene regulatory, gene co-expression, metabolic, signaling and neural networks. In PPI networks, proteins are the nodes, and the interactions between them are the edges (Han et al., 2004). These networks are evolutionarily conserved across species (Sharan et al., 2005). Their hub proteins are essential for the survival and function of the cell or organism. The overall structure of the network, not just the individual node pairs and edges, is crucial for the functioning of the cell or organism (Jeong, Mason, Barabasi, & Oltvai, 2001).

Reducing biological network complexity

Most complex networks are very dense in the number of edges (Breitkreutz et al., 2010). This hampers the visualization and the study of the properties of the network. For instance, clustering and other algorithms that assume that the network is sparse (number of nodes is similar to the number of edges), may not perform well (Mishra, Schreiber, Stanton, & Tarjan, 2007). Moreover, modeling and simulation of complex networks may be difficult mathematically and computationally (Walpole, Papin, & Peirce, 2013). To circumvent this, there is a need to reduce the complex networks into the central or core components that retain the main topological and dynamic structure (Zanudo & Albert, 2013).

Many methods have been developed to reduce the complexity of networks. Among them are topological centrality, essential gene set, coarse-graining and filtering approaches. The topological approaches identify and remove redundant links using network centrality measures (Newman, 2006). The essential gene set methods seek to determine the minimal gene set responsible for life sustenance (Commichau, Pietack, & Stulke, 2013; Kobayashi et al., 2003). Both the topological centrality and essential gene set techniques do not consider interactions between the major gene products and other vital signaling components.
Coarse-graining approaches reduce the complexity of networks by identifying network motifs, collapsing the network motif into a single node, and repeating the process until there are no motifs in the network (Itzkovitz et al., 2005; Song, Havlin, & Makse, 2005). The complexity of the resulting network is reduced. However, it loses the underlying network topology and weight distribution of the original network.

Network filtering identifies and retains significant links from the network's connectivity and its weight distribution using a null model (Grady, Thiemann, & Brockmann, 2012; Santoni, Pedicini, & Castiglione, 2008; Serrano, Boguna, & Vespignani, 2009; Tumminello, Aste, Di Matteo, & Mantegna, 2005). The network filtering algorithms are known to perform better in maintaining the complex structure of weighted networks, as well as their topology (Dianati, 2016; Grady et al., 2012; Radicchi, Ramasco, & Fortunato, 2011; Serrano et al., 2009; Tumminello et al., 2005). They are able to retain the multiscale structure inherent in natural complex networks. Each uses a null model to calculate the significance of the node (Serrano et al., 2009) or edge (Radicchi et al., 2011). The calculated p-value is then used to filter the nodes or edges, reducing the network into its central components. The methods are amenable to diverse networks, especially networks whose edges are weighted. Figure 1c is an example of a weighted network in which the strength of the interactions between the nodes are represented by weights.

Cell-type specific network model reconstruction

To reconstruct a biological network, data from observational, interventional, and perturbation experiments are usually integrated (Markowetz & Spang, 2007). The availability of high throughput technologies has enabled genome-scale reconstruction of signaling networks (Hyduke & Palsson, 2010). These large-scale networks, especially cell-type specific networks, are complex. Their reduction to the central or core components may lend them to systems biology modeling and simulation studies (Zanudo & Albert, 2013).

An enormous amount of PPI data is available in pathway and protein interaction data repositories (Klingström & Plewczynski, 2011). However, the data in these repositories are not cell-type specific. To reconstruct a cell-type specific network, it is possible to start with a PPI network of all proteins that could be expressed and function in the cell-type.

By integrating biological data available for the specific cell-type, its network can be reconstructed. Based on the purpose and scope of the reconstructed network, it can be refined with literature mining. The refined network can be used for network modeling and simulation, e.g. simulating the effects of genetic perturbations in a specific cell-type.

Overview of the immune system, variations and diseases

The immune system consists of a network of cells, tissues, and organs that identify, neutralize, destroy, and remove foreign pathogens from the body. The immune system is broadly classified as belonging to the innate and adaptive systems (Murphy et al., 2012). To mount a response, both innate and adaptive systems are capable of, first, differentiating between non-self from self antigens (Jiang & Chess, 2009; Medzhitov & Janeway, 2002), and second, mounting a response to neutralize and destroy the antigens or the invading cell (Iwasaki & Medzhitov, 2015).

The innate immune system

Cells of the innate immune system recognize and respond to invaders in a non-specific manner (Berg & Forman, 2006). The cells that perform adaptive response have evolved mechanisms that are unique to the type of non-self molecule or cell they respond to (Pancer & Cooper, 2006).

The innate immune system offers barriers against infectious or invading pathogens (Murphy et al., 2012). The barriers provided by the innate immune system include physical and mechanical (e.g. the skin and mucous membranes), chemical (e.g. lysozymes in saliva), and biological (e.g. the microbiome). The cells of the innate immune system include phagocytes that engulf and kill invading cells (the macrophages and neutrophils) and natural killer cells that kill invading organisms by secreting lethal chemicals (perforin and granzymes) into them.

The adaptive immune system

Unlike the innate immunity, the recognition and response to foreign antigens are very specific in adaptive immunity (Pancer & Cooper, 2006). The recognition of the adaptive immunity is so subtle that it is capable of distinguishing peptides that differ by a single amino acid. Since there are thousands of proteins and an even greater number of peptides derived from them, the adaptive immune system has evolved an enormous variability to recognize these peptides.

As mentioned above, the adaptive system must distinguish between self to non-self peptides (Jiang & Chess, 2009; Medzhitov & Janeway, 2002). Failure to do so leads to a range of diseases, from mild to lethal (Murphy et al., 2012). To distinguish self from non-self molecules, the adaptive immunity uses the negative selection of cells that bind strongly to self-peptides during their development in the primary lymphoid organs. This process prevents the immune system from attacking normal cells.

Defects in the negative selection mechanism is the root cause of many autoimmune disorders.

Adaptive immunity cells are activated by peptides called antigens. Antigens are presented on the surface of cells (Pancer & Cooper, 2006). Cells degrade proteins, process the peptides. The processed petides are then presented on the surface of the cell. Innate immune cells engulf and kill foreign microorganisms. The proteins from the microorganism are processed and presented as a peptide (epitope) bound to the major histocompatibility complex (MCH), on the surface of the cell. The adaptive immune cells are activated and elicit a response after binding and recognizing the epitope. This process is similar to non-innate immunity cells. Non-immunity cells degrade proteins, process them and present their fragments in a similar manner to the adaptive immune cells.

To ensure heightened immune response in future attacks by foreign antigens, the adaptive immune system confers immune memory after the initial attack (Kurtz, 2004). After the initial assault, the adaptive immune cells proliferate and remain in circulation for an extended period of time. Hence, during subsequent attacks, the response is heightened and swift.

In the adaptive system, antibodies are secreted to neutralize a specific antigen (Panda & Ding, 2015). The secreted antibodies are capable of binding and grouping many foreign antigens into a cluster. This makes it easy for other immune cells to attack, kill and eliminate the foreign antigens.

The T and B lymphocytes are the two primary groups of cells involved in adaptive immunity (Pancer & Cooper, 2006). Their initial development takes place in the primary lymphoid organs. They leave the primary lymphoid organs (bone marrow) into circulation and mature in the secondary lymphoid organs (spleen, lymph nodes, and others).

Lymphocytes: T cells and B cells

Lymphocytes are white blood cells whose development begins in the bone marrow and mature in the secondary lymphoid organs, including the thymus and the lymph nodes (Murphy et al., 2012). The lymphocytes are made of T cells and B cells, which are further differentiated into their subtypes according to their function, type of receptors and other molecular profiles.

Produced in the bone marrow, T cells mature in the thymus after undergoing negative selection (Murphy et al., 2012). They have a surface receptor, the T cell receptor (TCR), as well as coreceptors. They are categorized into two main subtypes, the helper T cells (Th) and the killer T cells. These two main groups are distinguished by the coreceptors to the TCR. The Th cells have CD4 as coreceptor, while the killer T cells have the CD8 coreceptor.

Th cells are the main regulators of the immune system (Swain et al., 1991). Antigens are presented to the TCRs of the Th cells by antigen presenting cells. If recognized, the T cell is activated, undergoes proliferation and secretes cytokines that activate B cells and various other immune system cells and pathways.

The killer T cells scout for foreign antigens by binding to the peptide:MHC I complexes on the surface of cells (Iwasaki & Medzhitov, 2015). When it binds and recognizes the peptide as non-self, it elicits a response that kills the antigen presenting cell.

Like T cells, B cells have the B cell receptor (BCR) and several coreceptors for antigen recognition and activation response (Kurosaki, Shinohara, & Baba, 2010). Activation-induced pathways are triggered when the BCR binds and recognizes the antigen presented by the MHC complex. The B cell gets activated when activated T cells release cytokines. The activated B cell proliferates through cloning into two cell populations, the plasma and memory cells.

Plasma cells release copious amounts of antibodies during infection (Murphy et al., 2012). The antibodies perform a range of functions. They bind foreign microorganisms enabling their destruction by the complement pathway or incapacitate them. They bind antigens enabling their elimination by macrophages. Antibodies also act as antitoxins and gather pathogens for easy elimination.

As mentioned above, the B and T memory cells have a prolonged life span and enable heightened response to secondary attacks (Kurtz, 2004). Since their response is specific to the antigen that caused the first infection, the subsequent response is swift and more potent.

The effects of variations on the immune system

Advancements in sequencing techniques have made genome sequencing and identification of variants cheaper and easier. In humans, 99.9% of the genome is identical, and the remaining 0.1% renders each genome unique. These variations are of a wide variety, from a single nucleotide substitution to chromosomal insertions, deletions or duplications. These variations are either inherited or occur in non-germinal cells. The most common genetic variation include single nucleotide variations (Altshuler et al., 2010).

Many inherited variations affect the immune system (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005). The effects of these variations range from mild to severe. Among the most studied of such variations are PIDs (Picard et al., 2015). Somatic variations are genetic alterations that occur in non-germinal cells after conception. These types of variations accumulate during a life time, generating between hundreds to thousands of genetic alterations in a healthy individual. Some

of these variations lead to AASs in proteins and cause diseases, especially cancer (Alexandrov et al., 2013).

Primary immunodeficiency

PIDs are genetic defects or variations in the immune system repertoire of genes and proteins, most of which are hereditary. These diseases are uncommon and present diverse and sometimes overlapping phenotypes. Due to the complex and diverse phenotypes, PIDs are difficult to diagnose. Prognosis of PIDs depends on early diagnosis. The severity of PIDs can range from benign to fatal (Samarghitean, Ortutay, & Vihinen, 2009). Thus, a lot of effort has been made to catalog and classify PIDs to facilitate early diagnosis, and hence, better prognosis. There are about 300 known PIDs (Picard et al., 2015; Piirilä, Väliaho, & Vihinen, 2006). By integrating these data, it is possible to perform global dynamic studies in the effect of PID-associated perturbations in the affected immune response cells.

Systems study of the effect of PIDs in T and B cells

Reconstructed interaction networks have been used to uncover the underlying mechanisms of biological processes in both normal and disease conditions (del Sol, Balling, Hood, & Galas, 2010; Goh et al., 2007). T and B cell protein interaction networks have been reconstructed to study the cellular dynamics of their activation and response (Chakraborty & Das, 2010). PID deficiencies disrupt essential biomolecular pathways for T and B cell activation and responses.

Several approaches are available to study the effects of such disruptions. Quantitative approaches require reaction constants for each interaction, most of which are not available (Aldridge, Burke, Lauffenburger, & Sorger, 2006). Moreover, quantitative methods also require the knowledge of kinetic reaction parameters that are hard to compute.

However, semi-quantitative and qualitative methods can be applied to larger networks (de Jong, 2002). These non-quantitative methods are able to characterize the dynamic trends of the system. Despite their wide application, semi-quantitative studies have not been used to investigate PID perturbations in major immune response cellular systems, like T cells and B cells.

Cancer immunogenicity

In cancers, there is a preponderance of somatic variations that result in nonsynonymous AASs. These AASs are the main drivers of many cancers. Processed proteins that contain the AASs can generate epitopes that T cells

recognize as non-self. These cancer-associated antigens are called neoantigens and can elicit an immune response against the cancer cells that process and present them.

Neoantigens can be used to treat cancer patients (Schumacher & Schreiber, 2015). Many studies have been conducted to investigate the effectiveness of neoantigens in cancer immunotherapy (Blankenstein, Leisegang, Uckert, & Schreiber, 2015; Schumacher & Hacohen, 2016; Tran, Robbins, & Rosenberg, 2017; Verdegaal et al., 2016; Vormehr et al., 2016). The most promising of these efforts have been in the development of cancer vaccines for personalized therapy (Desrichard, Snyder, & Chan, 2016). To identify cancer-associated neoepitopes, exome sequencing of both the normal and tumor tissues is performed (Schumacher & Schreiber, 2015).

However, it has been difficult to validate the neoepitopes as *bona fide* neoantigens (Anonymous, 2017). Thus, the identification of neoepitopes that are immunogenic is an ongoing research question, poised with great promises in personalized therapeutic applications. This is evident in the recent studies of neoantigen vaccine-based therapies that progressed to phase II of clinical trials with great success (Ott et al., 2017; Sahin et al., 2017).

Several tools have been developed to predict neoepitopes (Gfeller, Bassani-Sternberg, Schmidt, & Luescher, 2016). Most of these tools are developed for peptide-MHC I affinity. The performance of the tools varies according to data size and data composition (Kim et al., 2014; Trolle et al., 2015). Most of the methods depend on experimentally verified neoepitopes that are limited in size and diversity. Thus, the performance of the methods are adversely affected. As better assays are developed to experimentally verify predicted neoepitopes, and develop better prediction methods, neoepitope predictors might be improved. NetMHC, a tool for predicting MHC class I affinity to peptides, is one of the best methods in this category (Gfeller et al., 2016).

The occurrence and load of neoantigens have been performed in several studies, but details of the characteristics of neoantigens across cancer types are not available.

Disease diagnosis, therapy, and prognosis

Complications exist in the diagnosis and prognosis of diseases that either affect or evades the immune system. PIDs represent a large group of diseases that affect the immune system and present difficulties during diagnosis, which in turn affects prognosis. On the other hand, cancers are an example of complex diseases that evade the immune systems mechanisms. Due to the diagnostic and prognostic complications presented by these diseases, more efforts are needed to study the effects of variations to the immune response cell repertoire. In addition to studying the genotype-to-phenotype effects of individual PID proteins, their underlying

systemic effects to the cell can be investigated using systems biology methods. Characterizing the immunodominant features of neoantigens may also facilitate the advancement of vaccine-based therapies for cancers.

Research questions

The purpose of this thesis was to answer the following questions.

- How can the central components of a cell-type be identified with time series microarrays?
- How do the T cell PID proteins affect the T cell receptor-dependent activation dynamics?
- How do the B cell PID proteins affect the B cell receptor-dependent activation dynamics?
- What are the characteristics of neoepitopes analyzed from pan-cancer data?

Overview of methods

Protein-protein interaction network reconstruction

The immunome proteins were obtained from the Immunome Knowledge Base (IKB) and supplemented with those from relevant immune system pathways from KEGG data repository (Paper I) (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012). Next, the proteins were used to obtain experimentally verified and consolidated PPI data by retrieving PPIs for the immunome proteins from the iRefIndex database, version 9.0 (Razick, Magklaras, & Donaldson, 2008). The PPIs were filtered using ppiTrim, version 1.2.1 (Stojmirovic & Yu, 2011), an algorithm that maps protein interactants to NCBI gene identifiers and filters the PPIs as follows:

1. remove undesired raw interactions
2. deflate potentially expanded complexes, and
3. reconcile annotation labels from the different PPI databases.

Further filtering steps were performed by omitting the following PPIs:

1. non-experimentally verified PPIs
2. PPIs from experiments conducted on non-human cells or tissues
3. PPIs that are part of a complex
4. PPIs for which both interactants are from the same gene
5. multiple copies of binary PPIs, and
6. PPIs for which both interactants were not immunome proteins.

The data were analyzed with the R statistical programming environment (R-Core-Team, 2016). Network reconstruction and analysis were conducted with igraph library (Csardi & Nepusz, 2006) and network visualization with Cytoscape, version 2.8 (Kohl, Wiese, & Warscheid, 2011).

Gene expression data, preprocessing and analysis

Microarray datasets were obtained from the GEO (Sayers et al., 2012) and ArrayExpress (Parkinson et al., 2011) databases using the following criteria:

1. time course experiment
2. the experiment has ≥ 3 samples
3. the experiment has ≥ 1 sample as baseline, and
4. the experiment was conducted on the Affymetrix whole transcript array platform U133A, U133A 2.0, U133B, U133 plus 2.0 or U95A arrays, to reduce bias during data integration.

Pre-processing of the microarray datasets was performed with R/Bioconductor packages (Gentleman et al., 2004). The raw microarray datasets were pre-processed and quality controlled with box plots, arrayPLM and simpleaffy procedures (Wilson & Miller, 2005). The Robust Multi-Array method implemented in the affy package (Gautier, Cope, Bolstad, & Irizarry, 2004) was used for normalization. The gene expression values were obtained from their average probe set values. Further, expression data for the genes that did not code for the immunome proteins were removed.

Both ComBat and plotMDS are algorithms implemented in the inSilicoMerging package for integrating microarray datasets (Taminau et al., 2012). The preprocessed and normalized datasets obtained above were merged with ComBat. Batch effect and PCA analyses were performed with ComBat and plotMDS, respectively, to examine bias effects on the datasets.

The bootstrap package in R was used to calculate for all gene pair combinations, the average values of the jackknife Pearson correlation coefficient from the merged expression data. The absolute values of these correlation coefficients were used as weights for the immunome interactome links.

Protein network filtering

The immunome interactome network was reconstructed as an undirected and link-weighted graph with the igraph package in R. The immunome protein coding genes, the PPIs, and the correlation coefficients were denoted by nodes, links, and link weights, respectively. The GloSS algorithm (Radicchi et al., 2011), which filters a weighted network, retaining its core structure and weight distribution, was used to filter the immunome interactome network. GloSS uses a global null model to calculate link significance (p-value) by randomizing link weight assignment while

maintaining the structure of the network. The network links were filtered using the link p-values in decreasing order. Connectivity between the TCR complex and the NF-κB signaling pathways was tracked during the filtering process. The GloSS filtering procedure was as follows:

1. calculate link significance (p-value)
2. get least significant link (maximum p-value)
3. get rid of the least significant link
4. if a path exists between TCR and transcription factor, go to 2, and
5. if a path does not exist between TCR and transcription factor, return the link and stop filtering.

This procedure was performed for both the NF-κB and the NFAT signaling pathways. The largest shortest path of a network is its diameter. The region of a network where a path exists between all nodes is a connected component. Changes to the network diameter, the relative size of the largest connected component and the average size of the isolated components were tracked during the network filtering procedure. The relative size of the largest component is the number of nodes in the largest component divided by the number of nodes in the whole network. Igraph was used to plot the network scores against the fraction of filtered nodes represented by $(\text{No of deleted nodes})/(\text{No of nodes in the network})$. The network scores were calculated with the igraph package.

Robustness of the T cell PPI network

While keeping the topology unchanged, a proportion of the link-weights were randomized to obtain the weight-randomized networks. Link-weight randomized networks with 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 of links randomized were created. Thirty repetitions were performed on each link weight-randomized network as follows:

1. a proportion of links were randomly selected
2. their weights randomly reassigned
3. the filtering procedure performed
4. the nodes and the links measures for node degree, average path length, betweenness centrality, as well as clustering coefficient of the network, and the intersection between the TPPIN and the link weight-randomized networks, calculated, and
5. the mean network scores in step 4 were retained.

Gene Ontology term enrichment, over-representation, and semantic similarity analysis

GO term enrichment and semantic similarity were performed with WebGestalt (Zhang, Kirov, & Snoddy, 2005) and GOSemSim, version 1.18.0 R/Bioconductor package (Yu et al., 2010), respectively. The immunome interactome was the background for the GO enrichment analysis. The Fisher's exact test was calculated for significance. The Benjamini-Hochberg procedure was calculated for multiple comparisons. Semantic similarity between the immunome interactome and the TPPIN was also calculated.

Analysis of essential genes

From the Mouse Genome Informatics database (Drabkin, Blake, & Mouse Genome Informatics, 2012), human orthologs of lethality genes were retrieved. Genes with "wean" and "partial" lethality types were excluded. Non-immunome genes were removed, and significantly enriched genes in the TPPIN identified with the Fisher's exact test. The analysis and visualization were performed with the biomaRt package (Durinck, Spellman, Birney, & Huber, 2009) and Cytoscape, version 2.8, respectively.

Network reconstruction and analysis

In Paper II, after supplementing the TPPIN with literature mining, the naïve CD4+ T cell network model was reconstructed from Boolean equations that include central TCR/CD28 signaling components. In Paper III the B cell model was reconstructed directly from literature mining.

Data analysis, the interaction graph and network visualization were accomplished with the R software, the CellNetAnalyzer, version 2016.1 (Klamt, Saez-Rodriguez, & Gilles, 2007) and Cytoscape, version 3.3.0 (Demchak et al., 2014), respectively. The feedback and feedforward loops of the underlying interaction graph of the model were computed with NetDS, a Cytoscape, version 2.8 plugin (Le & Kwon, 2011). The igraph/R package was used to compute strongly connected components. A Boolean network model is represented by variables whose state or value is either 0 or 1. During simulation, a protein's state is calculated at each round or time step from the values of the influencing proteins, which are proteins that are connected to it. This procedure is performed at each time step for all proteins in the network.

In the Odefy software (Krumsiek, Poelsterl, Wittmann, & Theis, 2010) the Boolean update functions are converted to normalized HillCubes, a system of continuous ordinary differential equation (ODE), in which the proteins' states are in the range 0 to 1, inclusive (Wittmann et al., 2009). The parameters of the ODE system of equations include τ, $\overline{x}$, k and n that describe the life-time, decay and the activation at half-maximal level of the protein, and the cooperativity between the protein interactions, respectively.

Basin of attraction and attractor identification

The normalize HillCube update functions were used to simulate the dynamics of the naïve CD4+ T cell and the B cell models. Except for the parameter values in Table 1, default parameters were used. The default parameters were $n = 3$, $k = 0.5$ and $\tau = 1$. Each simulation returned both a basin of attraction and an attractor. The PID-perturbed attractors were obtained as follows:

1. convert the PID protein to an input node
2. change its state to reflect the perturbation as reported in the literature, and
3. while keeping the perturbed state of the PID protein unchanged and using the same parameters as for the wild-type simulation, perform normalized HillCube simulation until an attractor is reached.

Table 1. Tuned parameters of nodes in the Odefy-simulated T and B cell network model

Influenced node	Network model	Influencing node(s)	τ	n		k
PAG1	T cell	[]a	1	20		0.9
DAG	T cell	DGK	1	20		0.9
DGK	T cell	[]	1	20		0.9
DGK	T cell	[]	1	3		0.9
LCK	T cell	MAPK1	10	20		0.1
CBL	T cell	[]	3	20		0.9
CALN	T cell	CABIN1	1	3		0.9
CALN	T cell	RCAN1	1	3		0.9
CALN	T cell	AKAP5	1	3		0.9
LYN	B cell	DOK3	10	20		0.1
PAG1	B cell	PTPRC	20	32		0.9
CSK	B cell	PAG1	20	32		0.9
BCR	B cell	PTPN6	1	32		0.9
BCR	B cell	PTPN11	1	32		0.9
BCR	B cell	CSK	1	32		0.9
PIP2_2	B cell	INPP5D	16	32		0.9

[a]All influencing nodes. PAG1, phosphoprotein membrane anchor with glycosphingolipid microdomains 1; DAG, second messenger, diacylglycerol; DGK, diacylglycerol kinases; LCK, LCK proto-oncogene, Src family tyrosine kinase; MAPK1, mitogen-activated protein kinase 1 (ERK); CBL, Cbl proto-oncogene; CALN, calcineurin complex; CABIN1, calcineurin Binding Protein 1, RCAN1, regulator of calcineurin 1, AKAP5, A-kinase anchoring protein 5; PIP2_2, PtdIns(3,4)P2, Phosphatidylinositol 3,4-bisphosphate.

Primary immunodeficiency data

PID data were obtained from the IDbases (Piirilä et al., 2006), the International Union of Immunological Societies (IUIS) expert committee classification of PID data (Picard et al., 2015) and a review (Vihinen, 2015). The PIDs in Paper II included LCK, ZAP70, ITK, IKKB, NEMO, CARD11, MALT1, BCL10, NFKBIA, PTPRC, MAP3K14 and PI3K deficiencies, whereas those in Paper III were BCL10, CARD11, CD19, CD21, CD81, IKKB, KRAS, LYN, MALT1, MS4A1, NEMO, NFKB1, NFKBIA, ORAI1, PI3K, PLCG2, STIM1 and WIPF1 deficiencies.

Variation data

In Paper IV, the pan-cancer AASs were retrieved from (Alexandrov et al., 2013). Sequences for proteins coded by the genes in the pan-cancer dataset were obtained from Ensembl (Flicek et al., 2014). Twenty-one amino acid long wild-type and variant peptides were derived from the proteins such that the variant position was in the middle.

HLA-peptide binding affinity prediction

The peptide:HLA affinity predictions were performed with NetMHC 4.0 software (Andreatta & Nielsen, 2016). The peptide affinity predictions of $IC_{50} \leq 50$ nM, $50 < IC_{50} \leq 500$ nM and > 500nM denoted high, weak and non-binders, respectively. High binding variant peptides whose corresponding wild-type peptides were either weak or non-binders were defined as neoepitopes.

Data analysis of neoepitope enriched proteins

Data analyses were performed with R. The amino acid hydropathy and sequence logo analysis and visualization were done with the MultiDisp software (http://structure.bmc.lu.se/MultiDisp). The GO term enrichment analysis was achieved with GOrilla (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009). The summary and visualization of the GO term enrichment were conducted with REViGO (Supek, Bosnjak, Skunca, & Smuc, 2011).

Overview of results

Identifying core cell-specific protein interaction network

Immunome proteins

The T cell-specific PPI network was identified. To achieve this 1,579 of immune response proteins, of which 885 were from the IKB database (Paper I) (Ortutay & Vihinen, 2009b) and 694 were from the KEGG immune system pathways (Kanehisa et al., 2012), were obtained. These immunome proteins were used to generate the immunome interactome. PPI data were retrieved from the iRefIndex (Razick et al., 2008) and ppiTrim (Stojmirovic & Yu, 2011) was used for filtering. After the filtering, the PPIs retained were immunome protein-containing, experimentally verified, binary and non-redundant. The network consisted of 5,603 PPIs and 1,259 immunome proteins.

Immunome gene pair correlation data

16 time series datasets having 384 samples from 5 Affymetrix platforms were retrieved from the GEO (Sayers et al., 2012) and ArrayExpress (Parkinson et al., 2011) data repositories. These experiments were preprocessed and normalized. The normalized datasets were integrated after batch effects and quality control analysis. 1,149 of the 1,259 immunome protein coding genes, expressed in at least 80% of the samples were integrated and used for further analysis. The Jackknife Pearson correlation between all gene-pair combinations in the integrated dataset was computed. The correlations of 5,164 gene pairs for 1,140 genes coding for immunome proteins were obtained. The minimum, maximum and mean of the correlations were -0.06, 0.88 and 0.09, respectively.

Reconstructed immunome interactome and filtering

The T cell-specific PPI network was reconstructed with the immunome interactome as the nodes and links and the correlation data as link weights. Next, the network was filtered to identify its core without losing its underlying complex structure with the GloSS algorithm (Radicchi et al., 2011). Two aspects of the network were used to guide the filtering process. First, the biological information of the network that includes essential T cell pathways, the NF-κB and NFAT pathways, were used to inform the termination of the filtering procedure. Next, network topology measures that indicate network connectivity and robustness were used to monitor the the effect of the filtering process. These network measures include the network diameter, the relative size of the largest connected component and the average size of the isolated components. The filtered network was called the T cell PPI network (TPPIN).

Support for the core T cell PPI network

The TPPIN was supported by several empirical evidence. First, the distribution of the link weights was retained before and after the filtering procedure. This implies that the filtering process maintained the core network structure. Second, noise was introduced to the immunome interactome by randomly assigning different proportions of link weights, while maintaining the structure. The noise introduced to the immunome interactome reduced its connectivity, robustness, and integrity significantly, compared to the TPPIN. Third, GO term enrichment analysis (Gene Ontology, 2012), was performed using the TPPIN proteins. Most biological process terms were enriched with terms for T cell-specific functions, as well as for general immune response. Fourth, when the immunome interactome proteins were semantically compared to those of the TPPIN, significant overlap between both protein groups was found, showing that the TPPIN is both T cell-specific and immunome interactome representative. Fifth, essential genes enrichment analysis was performed with the TPPIN protein coding genes. The human orthologs of the mouse essential genes from the Mouse Genome Informatics database were used (Drabkin et al., 2012). Highly significant enrichment of essential genes of the TPPIN protein coding genes was found. Lastly, the TPPIN was mapped to crucial T cell activation and response pathways, including the TCR, JAK-STAT, MAPK pathways. TPPIN includes almost all the essential components of these pathways, including most components necessary for the early signaling events of the NF-κB, NFAT and AP1 transactivation pathways. Further, the mapping shows the interconnection between the pathways. These demonstrate that the filtering process was successful in capturing the core T cell-specific network.

Modeling and simulating PID perturbation effects on T and B cells

Reconstructing network models for naive T and B cells

The naïve T and B cell activation network models were reconstructed and used to simulate the semi-dynamic effects of PID perturbations (Papers II and III). To reconstruct the T cell network, the 227 TPPIN interactions were used as the basis for mining the literature. Eighty-five interactions that are crucial for T cell receptor (TCR) and CD28 coreceptor signaling were used to reconstruct the CD4+ T cell Boolean network model. Similarly, the literature was mined for the major interactions for the B cell receptor (BCR) activation and response. Boolean equations for the interactions were generated and used for further analysis. The Boolean equations were represented in the sum-of-product form (Klamt, Saez-Rodriguez, Lindquist, Simeoni, & Gilles, 2006). 19 and 20 nodes in the T cell and B cell model, respectively, were source nodes (i.e., had no incoming edges).

Underlying structure of the network models

The structure of the network models was probed for signaling paths from the receptors to downstream transcription factors. In the T cell model, paths from the TCR (signal 1) and coreceptor CD28 (signal 2) to major TCR-dependent activation transcription factors, NF-κB, NFAT and AP1 (Smith-Garvin, Koretzky, & Jordan, 2009) were examined. In the B cell model, paths from the BCR to major transcription factors ELK1, BCL6, EGR1, AP1, NFAT, and NF-κB were probed. To achieve this, the network models were converted into the underlying interaction graphs. The T cell interaction graph consisted of a connected component of 85 nodes and 146 links. Moreover, a strongly connected component was identified with 25 nodes and 48 links. This strongly connected component shows the part of the network where the signaling paths experience the most signaling cross-talk. The B cell interaction graph consisted of a single strongly connected component with 107 and 188 links, respectively.

Feedback loops (FFLs) and feedforward loops (FBLs) were used to identify proteins that are essential along signaling paths in the models. Proteins that are in many such loops are considered essential to the dynamics of the network. In both interaction graphs, the shortest loops had 2 nodes, while the longest had 20 nodes, for the T cell, and 27 for the B cell interaction graph, respectively. Most of the PID proteins were found in many loops in the T cell graph (LCK, 409; ZAP70,380; CBM, 316; CARD11, 312; BCL10, 210; ITK, 120; PI3K, 110 and MALT1 in 106 FBLs) and

in the B cell graph (BCL10, 724; BLNK, 912; BTK, 3952; CARD11, 2896; CD19, 1484; CD21, 371; CD81, 371; IKKB, 362; KRAS, 2394; LYN, 6216; NEMO, 1448; NFKB1 and NFKBIA, 2172; ORAI1, 2700; PI3K, 6840; PLCG2, 8208; STIM1, 4050). Unlike in the B cell graph, NEMO, IKKB, NFKBIA and MAP3K14 PID proteins were in none of the loops. The presence of the PID proteins in many loops shows that their disruption can significantly affect the normal signaling dynamics of the cells.

Simulating the wild-type scenarios

To make sure that the model agrees with the literature and can reproduce normal or wild-type TCR- and BCR-dependent activation, the network was refined with *in silico* validation. Normalized HillCube simulation (Krumsiek et al., 2010; Wittmann et al., 2009) was used to validate the network models. The state of nodes was iteratively changed while constraining major network components during simulations. Further, scenarios in which signal 1 and/or signal 2 are turned on or off, separately and together in each of the models, were investigated. When either signal was turned off, the transcription factors AP1 and NFAT, but not NF-κB, were activated for the T cell model. In the B cell model, except for NFAT and NF-κB, all other transcription factors (ATF2, BCL6, CREB1, EGR1, ELK1, ETS1, FOXO1, JUN, MEF2C) were turned on. When both signals 1 and 2 were turned on, the results were comparable to the literature (LeBien & Tedder, 2008; Mitchell, Vargas, & Hoffmann, 2016; Smith-Garvin et al., 2009). Simulations were performed until the network reached an attractor state.

With the T cell model, a cyclic attractor was obtained after 40 update cycles or arbitrary time points. The network stayed in the attractor state with a period of 20 arbitrary time points. With the B cell model, a two-phase simulation was performed. The first phase simulated the dynamics of the early activation cascades after the BCR and its coreceptors (CD19/21/81) were activated, while the second phase represented late BCR modulating signaling events. In the first phase, the B cell model reached a point attractor after 80 time points. Following this, the modulators of the BCR were turned on, and the simulation was executed until an attractor was reached again. This second attractor was reached after 230 time points. Like the T cell, the attractors of the B cell model were in accordance with the literature as seen in the activation of all the major activating transcription factor pathways.

Simulating PID perturbations

The models were used to study the semi-dynamic effects of PID perturbations. To achieve this, 12 and 22 T and B cell PID proteins were obtained, respectively, from the ImmunoDeficiency Resource (Samarghitean, Väliaho, & Vihinen, 2007), IDbases (Piirilä et al., 2006), the most recent classification by the IUIS expert committee for PIDs (Picard et al., 2015) and a recent review (Vihinen, 2015). Variations in these proteins affect T and B cells from the pre-CD4+ and pre-B cell developmental stages, respectively. The T cell PIDs included BCL10, CARD11, IKKB, ITK, LCK, MALT1, MAP3K14, NEMO, NFKBIA, PI3K, PTPRC, TRAC and ZAP70. The B cell PIDs included BCL10, BLNK, BTK, gain- and loss-of-function CARD11, CD19, CD21, CD40, CD81, IKKB, KRAS, LYN, MALT1, MS4A1, NEMO, NFKB1, NFKBIA, ORAI1, PI3K, PLCG2, PTPRC and STIM1 deficiencies. Normalized HillCube simulations for each PID perturbation had significant effects on all three major transactivation factor pathways. The exceptions were as follows:

1. the overexpression perturbations (PI3K and NFKBIA in the T cell model; CARD11, IKKB, KRAS, PI3K, and PLCG2 in the B cell model), and,
2. the knockout perturbations (BCL10, CARD11, MS4A1, PTPRC, WIPF1) in the B cell model.

Severity of PIDs

Next the severity of the PIDs was examined. Except for the knockin perturbed PIDs in both models, all other PIDs were associated to SCIDs and deficiencies, which are associated to infectious disease susceptibilities (van der Burg & Gennery, 2011). PID discovery is evolving and leads to the cataloging, classification, and prioritization to ease cheaper and earlier diagnosis (Ortutay & Vihinen, 2009a; Picard et al., 2015; Samarghitean et al., 2009). Perturbed simulations for proteins that are in many loops on the models were conducted to probe their dynamic effects. Like PIDs, most of the proteins had significant effects on the main signaling pathways. Interestingly, these proteins are also disrupted in most PID attractors. Further interrogation of these proteins with the Human Genome connectome (Itan & Casanova, 2015) showed that many are connected to PID proteins. Thus, these proteins could be investigated during gene prioritization. The proteins include ABL, LCK, MAPK1, PRKCQ, LAT, RAS and VAV1 in the T cell model, and IKKA, CRACR2A, GAB1, GRB2, and ITPR1 in the B cell model.

Characterizing neoepitopes: a pan-cancer analysis

A pan-cancer analysis was performed to characterize neoepitopes (Paper IV).

Sequence data and prediction

Data for 783,615 AASs on proteins experimentally identified in 30 cancer types were obtained and used in the analysis (Alexandrov et al., 2013). For each AAS, a 21-mer peptide was constructed with the AAS position at the center, for both the wild-type and the variant peptide, and used as the input to NetMHC. NetMHC made 4,706,079,200 affinity predictions between 8- to 11-mers, to 80 human HLAs.

Strong and weak peptide binders to HLAs

Strong binding peptide-HLA affinity was defined as those below 50 nM. Weak binders were those with affinity above 50 nM but below 500nM, and non-binders were defined as those with affinity above 500 nM.

The number of binders for wild-type and variant datasets was 41,667,139 and 44,853,374, respectively. The number of 9-mer binders was extremely larger than peptides of the other lengths.

Peptide binders

Less than 2% of all predictions were either weak or strong binders. The wild-type and variant datasets had similar distributions of AAS positions within the binding peptides. The distribution of binders in cancer types followed the individual cancer variation rate and was similar for both the wild-type and variant datasets. Thus, the overall distribution of binders was quite similar across cancers and also across HLAs, in both the wild-type and variant datasets.

Stong and weak binders

The proportion of wild-type and variant binders that are strong binders was 0.24% and 0.22%, and those for weak binders, 0.67%, and 0.72%, respectively, based on the total number of predictions. AAS positions within the binding peptides were distributed evenly, except for positions 10 and 11, both across the wild-type and variant datasets, and the weak and strong binders. Similar to the overall binders datasets, the distribution across cancer types and HLAs for wild-type and variant binders, were similar among weak and strong binders. Further, the proportion of 9-mers within HLAs was similar within weak and strong binders (like that for all

binders). Thus, the characteristics of the weak and strong binders were similar, when compared to each other, and to the overall peptide binders.

Neoepitopes

The characteristics of neoepitopes were subsequently studied. Neoepitopes were defined as strong binding peptides whose equivalent wild-type peptides had weak affinity or were predicted as non-binders. Of all the predicted peptides over 11 million (0.24%), from over 95% of all proteins fulfilled the criteria as neoepitopes. This implies that peptides derived from virtually every protein potentially generated antigenic epitopes.

Neoepitope distribution among n-mers and proteins

The 9-mers were the most abundant (72%) among neoepitopes, followed by 10-, 11-, and 8-mers. Less than 2% of variants gave rise to neoepitopes of all lengths, and only 6.1% of two different lengths. The proportion of neoepitopes generated by variants is in the range 1-231. Most variants generated only a few neoepitopes, while 0.17% generated many neoepitopes (1,282 variants yielded ≥100 neoepitopes).

Neoepitope distribution among HLAs, at AAS positions and cancer type

The distribution of neoepitopes across HLAs varied slightly, whereas there were considerable differences in the distribution of n-mers. The percentage of AASs at each amino acid position in the neoepitopes was very similar to those observed for the wild-type and variant binders. As observed for the wild-type and variant binders, the distribution of neoepitopes in the cancer types followed the cancer mutation rate, and the proportion of n-mers in each cancer type was almost uniform.

The neoepitope data were mapped to the patient data from which the AASs were derived to study the distribution of neoepitopes in the patients. The minimum, maximum and median number of neoepitopes per patient were 4, over a half a million and about 6,856, respectively. Cancer types with a high mutational burden also had the highest proportions of neoepitopes per patient.

GO term enrichment for proteins that yield many neoepitopes

GO analysis was performed to identify the functional group enrichment of the proteins from which neoepitopes are derived (Eden et al., 2009). All human proteins were used as background and the proteins from which neoepitopes are derived, were used as targets. The most enriched biological process term categories comprised nucleic acid and RNA metabolism, whereas nucleic acid and RNA binding

consisted of the most significant molecular function enriched terms. The most enriched cellular compartment terms included the nucleosome and the nucleus.

Neoepitope amino acid residue analysis

The MultiDisp software was used to investigate the effects of the AASs on neoepitopes by examining the frequency of the different types of amino acid residues at the AAS positions. Albeit similar residues in some positions, an enrichment of F, I, L, V, Y was observed, and a reduced frequency of D, E, R, S, and T residues at the last position of the neoepitopes compared to the wild-type peptides. The enriched amino acids were hydrophobic, which confirms results (Chowell et al., 2015). Similar amino acids were observed at other positions in both the neoepitopes and the wild-type.

Moreover, with the Kyte-Doolittle hydropathy scale, the hydropathic characteristics of the amino acid residues at the AAS positions in both the neoepitopes and the wild-type peptides were investigated. Hydropathy is an essential feature of epitopes that is connected to the binding preference of amino acids within the HLA binding sites. Some positions in HLAs are essential for recognition and response. However, our results show that hydrophobic characteristics are preferred in neoepitopes, at all sites and in all n-mers, than in wild-type peptides.

General discussion

The effects of variation in normal and disease conditions were described. The core T cell-specific network was reconstructed, in Paper I. In Papers II and III, the focus was on the effects of immunodeficiencies on naïve CD4+ T and B cells from the pre-T and pre-B cell developmental stages. In Paper IV, the effects of AASs to the preponderance of MHC I-associated neoepitopes, as well as their characteristics were investigated. Below, the results and their implications are discussed.

How can the central components of a cell-type be identified with time series microarrays?

The core T cell-specific PPI network was identified after filtering the immunome interactome (Paper I). A list of immunome proteins was curated from the IDR and the KEGG pathways database (Kanehisa et al., 2012; Samarghitean et al., 2007). The immunome interactome was constructed from the immunome protein set by obtaining and preprocessing PPIs maintained in the iRefIndex compendium of PPIs. To weigh the immunome interactome links, time series gene expression profiles for human T cells were obtained from public repositories, preprocessed, normalized and merged together after careful batch effect correction analysis. Using the merged data, the mean of the jackknife Pearson correlation between all gene pairs whose products were in the immunome protein set was used to weigh the links of the immunome interactome. The weighted immunome interactome network was filtered to obtain the TPPIN. The resulting TPPIN was investigated and supported with multiple sources of evidence, including GO term enrichment and semantic similarity.

The TCR activation and signaling is critical for T cell development (Smith-Garvin et al., 2009), survival and functions, and most important components involved in the early TCR signaling events, are present. Except for CD3G and CD3D, most components that participate in the activation of the TCR complex and ITAMs, the coreceptors (CD4 and CD8) and the Src family kinases (LCK and FYN) are present in the TPPIN. After activation of the TCR and the ITAMs, ZAP70 and other crucial adaptor proteins are activated, leading to the formation and stimulation of the

macromolecular signaling complex that leads to downstream events of the TCR activation pathways. Except for a few adaptor molecules, most components of the macromolecular complex are present in the TPPIN.

After the formation of the proximal molecular complex, PLCG1 is activated (Cruz-Orcutt, Vacaflores, Connolly, Bunnell, & Houtman, 2014). PLCG1 cleaves PIP2, forming DAG and IP3 as second messengers. DAG activates PRKCQ which in turn leads to the activation of the CARD11-BCL10-MALT1 (CBM) and the IKK complexes (Isakov & Altman, 2012; D. Wang et al., 2004). The activation of the IKK complex leads to the activation of NF-κB (Mitchell et al., 2016). Additionally, RASGRP is activated by DAG, which in turn activates the MAPK cascade of signaling events that culminates in the activation of FOS (Smith-Garvin et al., 2009). AP1, a transcription factor complex of FOS and JUN, is formed after the activation of JUN through a PRKCQ-dependent pathway (Liu, Shepherd, & Nelin, 2007). Further, IP3 activates CaN, which in turn activates the transcription factor NFAT through the calcium signaling pathway (Oh-hora & Rao, 2008). All three transcription factors, NF-κB, AP1, and NFAT, are crucial response factors of the TCR activation signaling (Smith-Garvin et al., 2009). Almost all signaling proteins essential for the downstream activation of the transcription factors are present in TPPIN.

T cell specific network studies have previously been done (Mendoza, 2006; Mendoza & Pardo, 2010; Mendoza & Xenarios, 2006; Saez-Rodriguez et al., 2007; R. S. Wang & Albert, 2011). Most of these studies are centered on transcriptional networks. Such networks are small, consisting of a few dozen well-known transcription factors and their targets, as nodes. In this study, using an unsupervised approach, the T cell specific network was identified, TPPIN consisting of 288 nodes and 227 links that were derived from the immunome interactome, composed of 1,149 nodes and 5,164 links. The size of TPPIN can be used for systems biology studies.

The knowledge of T cell biology was used as the criteria to stop the filtering procedure. Central to T cell biology are the TCR and downstream signaling pathways that lead to T cell activation. The transcription factors NF-κB and NFAT, though present in many cell-types, are crucial to T cell activation and function (Smith-Garvin et al., 2009). The connectivity between the TCR and NF-κB, as well as TCR and NFAT, were maintained during the filtering. This ensured that the remaining network retains the central signaling components relevant for T cells.

Functional annotation of the proteins in the TPPIN was investigated for independent lines of evidence for the T cell specificity of TPPIN. The enrichment of GO terms for the TPPIN proteins for both the biological process and molecular function categories were analyzed. The result showed highly significant terms that are central to T cell function. To investigate the similarity between the TPPIN and the

immunome interactome proteins, semantic similarity of GO terms in biological process and molecular function categories were analyzed. The results show highly significant similarity between the TPPIN and immunome interactome. Since the survival of a cell depends on a set of indispensable genes, the TPPIN was probed for the enrichment of these genes. The results showed a high significance in essential genes in the TPPIN. The above lines of evidence demonstrate the relevance of the filtering routine.

Most publicly available data for microarray gene expression profiles are of diverse designs. As a result, experimental datasets of diverse designs were used. However, preprocessing, normalization, and batch effect correction analysis were implemented to minimize the effects of bias in the data used for the correlation analysis.

Gene expression experiments measure the aggregate relative expression of the genes expressed in the tissue under study, and the coexpressed gene products are functionally related. In the same light, the gene expression correlation between gene pairs used as link weights in the immunome interactome provides the T cell specific data on the strength between the gene products at an aggregate level.

During the filtering process, TPPIN maintains most of its network integrity and connectivity. Network statistics that indicate the connectivity and robustness of a network were used to monitor the effect of link removal during the filtering. The results suggest that the connectivity and robustness of the TPPIN were maintained throughout the filtering process.

The main limitation of the filtering routine used to identify the TPPIN is in the availability of data. Time series expression profiles for the cell-type under investigation is required. Further, each experiment should have at least 3 samples, and a set of proteins have to be used to track the connectivity of the network and set a stopping criterion when central pathways are about to lose their connectivity and robustness. However, large amounts of high throughput experimental data are available to the public. Thus, the availability of data might not pose a big challenge in many cases.

How do the T cell PID proteins affect the T cell receptor-dependent activation dynamics?

In Paper II, the normalized HillCube method (Krumsiek et al., 2010) was used to study the effects of PID protein perturbations in a network model for naive CD4+ T cell reconstructed from literature mining and a previously published core T cell protein network (Paper I). With the normalized HillCube approach, the network was

refined, and *in silico* validated, and used to study the dynamic effects of PID protein perturbations. The model and simulation were able to reproduce the effects of knocking out PID proteins from naïve CD4+ T cells.

Attractors for knockout perturbations for LCK, PTPRC, TRAC, and ZAP70 caused the most severe effects compared to the wild-type. The attractors for BCL10, MALT1, CARD11, MAP3K14, NEMO and IKKB severely affected the NF-κB pathway. Knocking out any of these proteins may dysregulate the IKK complex, which prevents NFKBIA from proteasomal degradation. The the intact NFKBIA sequesters NFKB1 in the cytosol, preventing it from being transported into the nucleus for response gene transactivation (Mitchell et al., 2016). Minor effects were observed for perturbed knockouts of MAP3K14, NEMO, and IKKB. Knockin perturbations for PI3K and NFKBIA had no effect on the dynamics of the network. Most PID proteins were found in most of the FFLs. Our approach showed severe effects for perturbed proteins along non-redundant and core pathways. On the other hand, knockout perturbation of proteins located in the periphery or along pathways with redundant paths displayed minor effects.

Although no effect was observed for knockin perturbations, gain-of-function variations in the catalytic subunit of PI3K are linked to serious respiratory infections, cancer and T cell senescence (Angulo et al., 2013; Crank et al., 2014; Lucas et al., 2014). On the other hand, the heterozygous and truncated NFKBIA knock-in variants sequester NFKB1 in the cytosol, preventing it from transactivation of response genes (Courtois et al., 2003; Janssen et al., 2004; McDonald et al., 2007). Since no effect was observed in the overexpression or knockin perturbations for the NFKBIA and PI3K, more detailed quantitative dynamic simulations may be needed to study their effects.

Upon MHC-antigen binding of the TCR, the CD3 ITAMs are activated. This leads to a plethora of critical signaling events that culminates in the transactivation of response genes by the major TFs (Smith-Garvin et al., 2009). Therefore, perturbing the TCR causes severe dysregulation of signal transduction. A homozygous variant of the TCR causes TRAC deficiency with severe effects (Morgan et al., 2011). This is corroborated by the profound dysregulation of the attractor of TRAC knockout, in which all signaling pathways are blocked.

LCK takes part in activating the ITAMs after TCR ligation (Palacios & Weiss, 2004). LCK's catalytic activity is regulated positively by PTPRC (Thomas & Brown, 1999). Lack of LCK activity causes low numbers of, and unresponsive T cells, which causes infectious disease predisposition (Hauck et al., 2012; Palacios & Weiss, 2004; Sawabe et al., 2001). Thus, PTPRC deficiency, caused by several variants in different individuals, including large deletions and AASs, causes severe phenotypic effects (Cale et al., 1997; Kung et al., 2000; Tchilian et al., 2001). This

is confirmed by disrupted signaling in all activation response pathways in the attractor of TRAC deficiency.

The absence of LCK disrupts NFAT and abrogates T cell response, which in turn leads to defects in TCR, NF-κB and calcium signaling (Hauck et al., 2012). LCK PID is associated with many disease phenotypes, including CD4+ T cell lymphopenia and respiratory tract infections. Our simulations confirmed the abrogation of the NF-κB, calcium signaling, and thus NFAT and AP1 pathways in the LCK PID-perturbed attractor.

Due to the proximity between ZAP70 and LCK in the early events of TCR signaling their knockout effects are similar. Many ZAP70 deficient patients have been diagnosed with severe disease phenotypes (Karaca et al., 2013; Picard et al., 2009; Schroeder, Triggs-Raine, & Zelinski, 2016). The ZAP70 knockout attractor shows significantly impaired signaling for major downstream effectors, especially in the calcium signaling pathways.

ITK is a major component of the LAT signalosome and TCR proximal signal transduction (Malissen, Aguado, & Malissen, 2005). The ITK deficiency is caused by several heterozygous variants that are associated with many disease phenotypes, including naïve CD4+ lymphopenia and recurrent infections (Ghosh, Bienemann, Boztug, & Borkhardt, 2014; Huck et al., 2009; Linka et al., 2012; P. Stepensky et al., 2011). Genotypic studies also associate this PID with several signaling defects, including activation-induced cell death and defective TCR activation signals. In these simulations, pathways downstream of the LAT signalosome are disrupted for all three major transcription factors. This is in agreement with defective T cell numbers and defective TCR activation and response.

After the formation of the LAT signalosome, PRKCQ is recruited and activated (Isakov & Altman, 2012). PRKCQ activates CARD11, which in turn binds and activates BCL10 and MALT1 to form the CBM complex (Thome, 2004; D. Wang et al., 2004). The PIDs associated with the components of the CBM complex are caused by homozygous variations in their genes (Jabara et al., 2013; Polina Stepensky et al., 2013; Torres et al., 2014). These PIDs cause diseases of diverse phenotypes in diagnosed patients. The CBM complex is an essential signaling component in the NF-κB pathway (Mitchell et al., 2016). CBM complex deficiencies are related to many T cell defects, including predominantly naïve CD4+ T cells and defective NF-kB pathway signaling. Our simulations confirm these findings as seen in the abrogated NF-κB and AP1 pathways in the attractors of the CARD11, BCL10, and MALT1.

The IKK complex, a major regulator of NF-κB, consists of the kinases IKKA and IKKB and NEMO, a regulatory protein (Smith-Garvin et al., 2009). The CBM complex activates TRAF6 through polyubiquitination, which in turn activates

MAP3K7 (Turvey et al., 2014). MAP3K7 regulates the assembly of the IKK complex. IKKB deficiency is associated with homozygous, duplicating and nonsynonymous substitution variants that cause many life-threatening diseases. Although IKKB deficiencies are connected with normal T cell numbers, the T cell subsets are low, peripheral T cells are non-responsive, and NF-κB signaling is disrupted (Pannicke et al., 2013). NEMO deficiencies are caused by AAS and exon skipping variations and are connected to several diseases, including colitis and ectodermal dysplasia (Fusco et al., 2015). Like IKKB, NEMO deficiencies are associated with normal T cell count but impaired TCR and NF-κB signaling. The attractors for IKKB and NEMO show severely impaired NF-κB signaling, while AP1 and NFAT signaling had minor effect, confirming the previous findings.

MAP3K14 is a major component in both the canonical and noncanonical NF-κB pathways (Mitchell et al., 2016). In the canonical pathway, MAP3K14 is activated via the AKT1 pathway and is involved in the activation-induced degradation of NFKBIA. In the noncanonical pathway, activated MAP3K14 associates with IKKA. This association mediates the degradation of the p100 unit of the NF-κB complex, allowing the nuclear transport of the NFKB2 dimers and transactivation response. The MAP3K14 deficiencies are caused by variants at its kinase activity site (Willmann et al., 2014). This causes impairment of both the canonical and non-canonical NF-κB signaling pathways and leads to severe microbial infections. This PID is associated with normal T cell numbers but inadequate activation response. In the MAP3K14 perturbed attractor, AP1 and NFAT pathways are unaffected while the NF-κB pathway was abrogated.

The above results capture the trends in the dynamic effects of knocked out PID proteins. Generally, more severe defects are associated with PIDs that are involved in the early events of the TCR signal transduction, while downstream events are less severe unless the perturbation is along a non-redundant signaling path to a crucial transcription effector. This work is the first to my knowledge that investigates the dynamic effects of PID proteins in CD4+ T cells using a network model.

Several proteins occurred along many loops in the network, and were found to be necessary for TCR response, were dysregulated in several PID perturbed attractors, and connected with disease phenotypes. Further, most of these proteins have been proposed as candidates during PID diagnosis, and are highly connected to PID proteins in the Human Gene Connectome (Itan & Casanova, 2015). Thirteen of the proteins are kinases, and 3 have guanyl-nucleotide exchange factor activity. Except for 7 genes, all are linked to diseases. Although these proteins have not yet been associated with any PID, they are strong candidates to be considered during diagnosis.

Several studies use diverse methods to arrive at a small set of suggested candidates for PID diagnosis (Itan & Casanova, 2015; Keerthikumar et al., 2009; Ortutay &

Vihinen, 2009a). Our approach which accounts for the dynamic signaling effects that PID perturbation has on the T cell network model, coupled with several sources of evidence, permitted us to detect the proposed set of proteins as candidate PIDs.

How do the B cell PID proteins affect the B cell receptor-dependent activation dynamics?

The literature was mined and the naïve B cell network model reconstructed and used it to investigate the dynamic effects of PID perturbations (Paper III). The reconstructed and refined network was *in silico* validated and used to simulate its dynamics. The normalized HillCube approach (Krumsiek et al., 2010) was used for simulating the wild-type, as well as the PID perturbations of the network. The results both recapitulates previous studies, and reveal novel dynamic effects of PID-dependent failure modes.

Profound defects were observed in the LYN, BTK, STIM1, ORAI1, CD19, CD21, and CD81 perturbed attractors compared to the wild-type. These are crucial components of BCR-dependent B cell activation (Dal Porto et al., 2004; LeBien & Tedder, 2008), and is captured by the severe defects trend in the perturbed attractors. On the other hand, lesser effects were observed in the perturbed attractors for BCL10, IKKB, loss-of-function CARD11, MALT1, NEMO and WIPF1 deficiencies.

The BCR-dependent B cell activation signaling culminates in the transcription of response genes by major transcription factors, each of which is necessary for the cell's function (Dal Porto et al., 2004; LeBien & Tedder, 2008). Receptor signals are transduced through adaptors and effectors to the downstream transcriptional regulators. Perturbing these signaling pathways may cause slight to complete impairment in the response controlled by the affected transcription factor. The PID attractors for MALT1, CARD11, PI3K and PLCG2 knockouts disrupted the NF-κB pathways. BCL6, a transcription factor that suppresses apoptotic and DNA damage signals, whose activity is reduced during BCR activation (Basso & Dalla-Favera, 2012; Basso et al., 2005), is turned off in all perturbed attractors. The pathways for EGR1, ELK1, and ETS1, transcriptional controllers for survival, proliferation and differentiation (Healy et al., 1997; Sementchenko & Watson, 2000; Yasuda et al., 2008), are dysregulated in several PID attractors, including those for BTK, CD19 and STIM1. The remaining transcription factors that are involved in controlling survival and proliferation in BCR-dependent B cell activation are disrupted in at least a few of the PID perturbed attractors.

The attractors for the knockin or overexpression PID perturbations, including PI3K, gain-of-function CARD11, KRAS, and NFKBIA, showed no effect compared to the wild-type. Hence, comprehensive kinetic approaches are required to study the consequences of these overexpression PIDs in the naïve B cell network dynamics.

All the PID proteins were found along loops. However, LYN, STIM1, ORAI1, and CD19 were found in most of the loops. Interestingly, LYN, STIM1, ORAI1 and CD19 perturbation had the most severe effects in the simulations. This is expected as the dynamics of a signaling network is closely related to the structure of its cycles or loops.

What are the characteristics of neoepitopes analyzed from pan-cancer data?

Understanding the features that make epitopes antigenic will facilitate cancer diagnosis and therapies (Capietto, Jhunjhunwala, & Delamarre, 2017; Schumacher & Schreiber, 2015). In Paper IV the features of neoepitopes across 30 cancer types were studied. A preponderance of neoepitopes was observed in all cancers. The predictions were made with NetMHC, a peptide-MHC class I affinity predictor, which is among the best performing predictors (Andreatta & Nielsen, 2016; Gfeller et al., 2016). However, the software overpredicts neoepitopes, and a huge number of wild-type peptides were predicted as neoepitopes. This is likely not possible because the immune system uses negative selection mechanisms to maintain self-tolerance and avoid autoimmune diseases.

The most likely neoepitopes were defined as high binding variant peptides whose corresponding wild-type either binds weakly or not at all. This yielded over 11 million neoepitopes, constituting 0.24% of the studied peptides. Although amino acids were uniformly distributed to peptide positions and that peptides of different lengths were similarly distributed among HLAs, 9-mers were the most abundant, accounting for close to three-quarters of the neoepitopes.

95.44% of the proteins in this study were retained after filtering for most likely neoepitopes. It is improbable that all the neoepitopes are antigenic. The biological mechanisms that lead to neoantigens for T cells are complex and not yet fully understood. Thus, prediction tools may over-predict, since their algorithms do not incorporate the full biological mechanism.

Peptide binding to an HLA, the peptide-HLA complex ligation to an antibody, or receptor that leads to an adaptive immune response depends on several biological processes. Further, the protein from which the peptide is derived is degraded in the cytosol and the peptides are transported and processed in the endoplasmic reticulum.

The processed peptide binds to the HLA and is presented and binds to antibodies or receptors to elicit an immune response. To keep ATP expenditure to the minimum, only peptides with high binding affinity are processed and presented. Besides high affinity, the amount and stability of the peptide is also essential for its antigenicity. Moreover, because the peptides have to be recognized as non-self to elicit an immune response, only a small fraction of predicted neoepitopes are immunodominant.

The number of experimentally verified neoantigens is limited. A previous study with vaccinia virus indicates that only a tiny fraction of peptides bind HLAs with high affinity (Assarsson et al., 2007). However, due to a large number of possible peptides under consideration, the number of potential antigenic peptides is in the order of thousands.

Although neoepitopes and their usage in clinical applications have been discussed (Boegel, Lower, Bukur, Sahin, & Castle, 2014; Brown et al., 2014; Hartmaier et al., 2017; Linnemann et al., 2015; Matsushita et al., 2016; Pritchard et al., 2015), there are also notes that immunogenic neoepitopes are uncommon (Anonymous, 2017). Correlating the findings in this study to the knowledge on T-cell response to the vaccinia virus WR strain (Moutaftsi et al., 2006), the number of effective peptides in the dataset from this study will be about 1.3 million peptides, which is still a large number.

Although there are thousands of HLA alleles, an individual has six HLA genes. Databases for HLA allele information include 12,351 class I alleles (Robinson et al., 2015). The most common alleles are very recurrent. Thus the results reported here are representative of human populations.

1,282 variants yield more than 100 peptides of various lengths that are neoepitopes to diverse HLAs. It is likely that peptides like these can raise T cell response, and hence, would be an important set to consider for therapy and other applications.

The GO term analysis showed enrichment of nucleic acid metabolism and RNA metabolic processes as the most significant biological process terms. Molecular functions of neoepitopes included terms for nucleic acid binding, protein dimerization, receptor binding, catalytic and protein complex binding. Additionally, several cellular compartments were enriched.

The most frequent variants originate from proteins that have catalytic, transporter and binding activity. Among proteins with the largest number of neoepitopes, only a few are known cancer proteins. However, almost all the proteins in the Cancer Gene Consensus yield numerous neoepitopes.

Conclusions

The filtering routine used to identify the TPPIN can be used to retain the core cell-type specific PPI network and can be applied for any cell-type for which sufficient time series gene expression datasets exist. This provides a means to study, model and simulate protein networks for any cellular system, however complex its network may be.

Diagnosis and prognosis of PIDs are frequently challenging. A novel approach was provided to investigate the effects of PIDs on naïve CD4+ T cell and B cell signal transduction network models. Novel proteins that may be investigated during diagnosis were also highlighted. This method is applicable in studying the effect of PIDs of any cellular system, including non-immune system diseases.

Most tools that perform well in predicting peptide-HLA binding affinity will improve in performance with the availability of more experimentally verified neoantigen data. With improved predictions and better experimental assays, neoantigen applications for diagnosis and cancer therapy will become personalized and established in clinical settings. The characteristics of neoepitopes presented in this study can be used to inform cancer vaccine development and potentially reduce its cost.

Acknowledgements

I would like to express sincere gratitude to my supervisor, Prof. Mauno Vihinen, for the many years of education I have had under his supervision, for his patience, kindness, and understanding during difficult times, for his encouragement, mentorship, and inspiration. This thesis would not have been possible without him.

My sincere thanks go to Prof. Bairong Shen for the opportunity he gave me to visit and work in his lab. I also want to thank all the students and staff at Bairong's lab, and Yang Yang, for the kind support and assistance I got during the visit.

To Csaba Ortutay, for the close support and supervision you gave me, the advice and open discussions, I thank you. Many thanks to my former lab mates in BMT Tampere, especially Jukka, for your advice and support, thank you very much.

My sincere gratitude to past and present lab colleagues. Many thanks to, Abhishek Niroula, for his kind support. Those desperate moments, those immensely joyful moments, the waves of laughter, we shared together, were, and remain exceptional for the time I spent working on this thesis. Thank you very much for the wonderful experience. I also want to thank Gerard for his assistance and support, including proofreading this thesis.

Many thanks to Barncancerfonden and the Faculty of Medicine, Lund University for funding this research project. Special thanks to Pierre, Sebastian, Martin, and Rikard for their IT support.

My immense gratitude to all my friends in Lund, for their support and encouragements. Special thanks to CamLund, for providing me with an extended family, support and above all, making life easier and fun during these few years. I am grateful to Sima, for being my mentor; to Said for his wise discussions and advice, to Mahnaz, for her support and assistance.

Words cannot describe my gratitude to my wife and children, Dibo, Oben, and Naseem, for their unconditional love, support, encouragement, understanding, care, and immense sacrifice. This thesis would not have been possible without them. Without the love, care and nurturing of my late parents, I will not be here today. I will forever be indebted to them for all they sacrificed for me. For my siblings, special gratitude for always being there in times of help and need. Special gratitude to Maitre Colins, for the love, care and support and advice. I thank very much.

Many thanks to all who have supported and encouraged me to make this thesis a success. Special thanks to Sahba and Tambe for proofreading this thesis, Nachida, John, Bright, Mushu, Remi, Raymond, Ivo, Jasper, Delphine(s), Angelbert, Divine, Minette, Joe, Ben and family, Remi, Eli and family, for your constant moral support.

References

Albert, R., & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47-97. doi:DOI 10.1103/RevModPhys.74.47

Aldous, J. M., & Wilson, R. J. (2000). *Graphs and applications : an introductory approach*: London : Springer, cop. 2000.

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., & Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nature Cell Biology, 8*(11), 1195-1203. doi:10.1038/ncb1497

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., . . . Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature, 500*(7463), 415-421. doi:10.1038/nature12477

Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., . . . Consortium, G. P. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi:10.1038/nature09534

Andreatta, M., & Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics, 32*(4), 511-517. doi:10.1093/bioinformatics/btv639

Angulo, I., Vadas, O., Garcon, F., Banham-Hall, E., Plagnol, V., Leahy, T. R., . . . Nejentsev, S. (2013). Phosphoinositide 3-kinase delta gene mutation predisposes to respiratory infection and airway damage. *Science, 342*(6160), 866-871. doi:10.1126/science.1243292

Anonymous. (2017). The problem with neoantigen prediction. *Nature Biotechnology, 35*(2), 97. doi:10.1038/nbt.3800

Assarsson, E., Sidney, J., Oseroff, C., Pasquetto, V., Bui, H. H., Frahm, N., . . . Sette, A. (2007). A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *Journal of Immunology, 178*(12), 7890-7901.

Barabasi, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews: Genetics, 12*(1), 56-68. doi:10.1038/nrg2918

Basso, K., & Dalla-Favera, R. (2012). Roles of BCL6 in normal and transformed germinal center B cells. *Immunological Reviews, 247*(1), 172-183. doi:10.1111/j.1600-065X.2012.01112.x

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature Genetics, 37*(4), 382-390. doi:10.1038/ng1532

Berg, R. E., & Forman, J. (2006). The role of CD8 T cells in innate immunity and in antigen non-specific protection. *Current Opinion in Immunology, 18*(3), 338-343. doi:10.1016/j.coi.2006.03.010

Blankenstein, T., Leisegang, M., Uckert, W., & Schreiber, H. (2015). Targeting cancer-specific mutations by T cell receptor gene therapy. *Current Opinion in Immunology, 33*, 112-119. doi:10.1016/j.coi.2015.02.005

Boegel, S., Lower, M., Bukur, T., Sahin, U., & Castle, J. C. (2014). A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *Oncoimmunology, 3*(8). doi:10.4161/21624011.2014.954893

Bollobás, B. (2001). *Random graphs*. Cambridge: Cambridge Univ. Press.

Braun, P., Carvunis, A. R., Charloteaux, B., Dreze, M., Ecker, J. R., Hill, D. E., . . . Co, A. I. M. (2011). Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science, 333*(6042), 601-607. doi:10.1126/science.1203877

Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L., Neduva, V., Larsen, B., . . . Tyers, M. (2010). A Global Protein Kinase and Phosphatase Interaction Network in Yeast. *Science, 328*(5981), 1043-1046. doi:10.1126/science.1176495

Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., & Holt, R. A. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Research, 24*(5), 743-750. doi:10.1101/gr.165985.113

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*: Oxford : Oxford University Press, 2007.

Cale, C. M., Klein, N. J., Novelli, V., Veys, P., Jones, A. M., & Morgan, G. (1997). Severe combined immunodeficiency with abnormalities in expression of the common leucocyte antigen, CD45. *Archives of Disease in Childhood, 76*(2), 163-164.

Capietto, A. H., Jhunjhunwala, S., & Delamarre, L. (2017). Characterizing neoantigens for personalized cancer immunotherapy. *Current Opinion in Immunology, 46*, 58-65. doi:10.1016/j.coi.2017.04.007

Chakraborty, A. K., & Das, J. (2010). Pairing computation with experimentation: a powerful coupling for understanding T cell signalling. *Nature Reviews: Immunology, 10*(1), 59-71. doi:10.1038/nri2688

Chowell, D., Krishna, S., Becker, P. D., Cocita, C., Shu, J., Tan, X., . . . Anderson, K. S. (2015). TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proceedings of the National Academy of Sciences of the United States of America, 112*(14), E1754-1762. doi:10.1073/pnas.1500973112

Commichau, F. M., Pietack, N., & Stulke, J. (2013). Essential genes in Bacillus subtilis: a re-evaluation after ten years. *Molecular Biosystems, 9*(6), 1068-1075. doi:10.1039/c3mb25595f

Courtois, G., Smahi, A., Reichenbach, J., Doffinger, R., Cancrini, C., Bonnet, M., . . . Casanova, J. L. (2003). A hypermorphic IkappaBalpha mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. *Journal of Clinical Investigation, 112*(7), 1108-1115. doi:10.1172/JCI18714

Crank, M. C., Grossman, J. K., Moir, S., Pittaluga, S., Buckner, C. M., Kardava, L., . . . Rosenzweig, S. D. (2014). Mutations in PIK3CD can cause hyper IgM syndrome

(HIGM) associated with increased cancer susceptibility. *Journal of Clinical Immunology, 34*(3), 272-276. doi:10.1007/s10875-014-0012-9

Cruz-Orcutt, N., Vacaflores, A., Connolly, S. F., Bunnell, S. C., & Houtman, J. C. D. (2014). Activated PLC-gamma 1 is catalytically induced at LAT but activated PLC-gamma 1 is localized at both LAT- and TCR-containing complexes. *Cellular Signalling, 26*(4), 797-805. doi:10.1016/j.cellsig.2013.12.022

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

Dal Porto, J. M., Gauld, S. B., Merrell, K. T., Mills, D., Pugh-Bernard, A. E., & Cambier, J. (2004). B cell antigen receptor signaling 101. *Molecular Immunology, 41*(6-7), 599-613. doi:10.1016/j.molimm.2004.04.008

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology, 9*(1), 67-103. doi:10.1089/10665270252833208

de Silva, E., & Stumpf, M. P. (2005). Complex networks and simple models in biology. *J R Soc Interface, 2*(5), 419-430. doi:10.1098/rsif.2005.0067

del Sol, A., Balling, R., Hood, L., & Galas, D. (2010). Diseases as network perturbations. *Current Opinion in Biotechnology, 21*(4), 566-571.

Demchak, B., Hull, T., Reich, M., Liefeld, T., Smoot, M., Ideker, T., & Mesirov, J. P. (2014). Cytoscape: the network visualization tool for GenomeSpace workflows. *F1000Res, 3*, 151. doi:10.12688/f1000research.4492.2

Desrichard, A., Snyder, A., & Chan, T. A. (2016). Cancer Neoantigens and Applications for Immunotherapy. *Clinical Cancer Research, 22*(4), 807-812. doi:10.1158/1078-0432.CCR-14-3175

Dianati, N. (2016). Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Physical Review E, 93*(1). doi:10.1103/PhysRevE.93.012304

Drabkin, H. J., Blake, J. A., & Mouse Genome Informatics, D. (2012). Manual Gene Ontology annotation workflow at the Mouse Genome Informatics Database. *Database: The Journal of Biological Databases and Curation, 2012*, bas045. doi:10.1093/database/bas045

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(8), 1184-1191. doi:10.1038/nprot.2009.97

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics, 10*, 48. doi:10.1186/1471-2105-10-48

Erdos, P., & Renyi, A. (1960). On the Evolution of Random Graphs. *Bulletin of the International Statistical Institute, 38*(4), 343-347.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Searle, S. M. (2014). Ensembl 2014. *Nucleic Acids Research, 42*(Database issue), D749-755. doi:10.1093/nar/gkt1196

Fusco, F., Pescatore, A., Conte, M. I., Mirabelli, P., Paciolla, M., Esposito, E., . . . Ursini, M. V. (2015). EDA-ID and IP, two faces of the same coin: how the same

IKBKG/NEMO mutation affecting the NF-kappaB pathway can cause immunodeficiency and/or inflammation. *International Reviews of Immunology, 34*(6), 445-459. doi:10.3109/08830185.2015.1055331

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics, 20*(3), 307-315. doi:10.1093/bioinformatics/btg405

Gene Ontology, C. (2012). The Gene Ontology: enhancements for 2011. *Nucleic Acids Research, 40*(Database issue), D559-564. doi:10.1093/nar/gkr1028

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80. doi:10.1186/gb-2004-5-10-r80

Gfeller, D., Bassani-Sternberg, M., Schmidt, J., & Luescher, I. F. (2016). Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology, 5*(7), e1177691. doi:10.1080/2162402X.2016.1177691

Ghosh, S., Bienemann, K., Boztug, K., & Borkhardt, A. (2014). Interleukin-2-inducible T-cell kinase (ITK) deficiency - clinical and molecular aspects. *Journal of Clinical Immunology, 34*(8), 892-899. doi:10.1007/s10875-014-0110-8

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America, 104*(21), 8685-8690. doi:10.1073/pnas.0701361104

Grady, D., Thiemann, C., & Brockmann, D. (2012). Robust classification of salient links in complex networks. *Nat Commun, 3*, 864. doi:10.1038/ncomms1847

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research, 33*, D514-D517.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., . . . Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature, 430*(6995), 88-93. doi:10.1038/nature02555

Hartmaier, R. J., Charo, J., Fabrizio, D., Goldberg, M. E., Albacker, L. A., Pao, W., & Chmielecki, J. (2017). Genomic analysis of 63,220 tumors reveals insights into tumor uniqueness and targeted cancer immunotherapy strategies. *Genome Medicine, 9*(1), 16. doi:10.1186/s13073-017-0408-2

Hauck, F., Randriamampita, C., Martin, E., Gerart, S., Lambert, N., Lim, A., . . . Picard, C. (2012). Primary T-cell immunodeficiency with immunodysregulation caused by autosomal recessive LCK deficiency. *Journal of Allergy and Clinical Immunology, 130*(5), 1144-1152 e1111. doi:10.1016/j.jaci.2012.07.029

Healy, J. I., Dolmetsch, R. E., Timmerman, L. A., Cyster, J. G., Thomas, M. L., Crabtree, G. R., . . . Goodnow, C. C. (1997). Different nuclear signals are activated by the B cell receptor during positive versus negative signaling. *Immunity, 6*(4), 419-428.

Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 79*(8), 2554-2558. doi:DOI 10.1073/pnas.79.8.2554

Huck, K., Feyen, O., Niehues, T., Ruschendorf, F., Hubner, N., Laws, H. J., . . . Borkhardt, A. (2009). Girls homozygous for an IL-2-inducible T cell kinase mutation that leads to protein deficiency develop fatal EBV-associated lymphoproliferation. *Journal of Clinical Investigation, 119*(5), 1350-1358.

Hyduke, D. R., & Palsson, B. O. (2010). Towards genome-scale signalling network reconstructions. *Nature Reviews: Genetics, 11*(4), 297-307. doi:10.1038/nrg2750

Ings, T. C., Montoya, J. M., Bascompte, J., Bluthgen, N., Brown, L., Dormann, C. F., . . . Woodward, G. (2009). Ecological networks - beyond food webs. *Journal of Animal Ecology, 78*(1), 253-269. doi:10.1111/j.1365-2656.2008.01460.x

Isakov, N., & Altman, A. (2012). PKC-theta-mediated signal delivery from the TCR/CD28 surface receptors. *Frontiers in Immunology, 3*, 273. doi:10.3389/fimmu.2012.00273

Itan, Y., & Casanova, J. L. (2015). Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Frontiers in Immunology, 6*, 142. doi:10.3389/fimmu.2015.00142

Itzkovitz, S., Levitt, R., Kashtan, N., Milo, R., Itzkovitz, M., & Alon, U. (2005). Coarse-graining and self-dissimilarity of complex networks. *Physical Review E, 71*(1). doi:10.1103/PhysRevE.71.016127

Iwasaki, A., & Medzhitov, R. (2015). Control of adaptive immunity by the innate immune system. *Nature Immunology, 16*(4), 343-353. doi:10.1038/ni.3123

Jabara, H. H., Ohsumi, T., Chou, J., Massaad, M. J., Benson, H., Megarbane, A., . . . Geha, R. S. (2013). A homozygous mucosa-associated lymphoid tissue 1 (MALT1) mutation in a family with combined immunodeficiency. *Journal of Allergy and Clinical Immunology, 132*(1), 151-158. doi:10.1016/j.jaci.2013.04.047

Janssen, R., van Wengen, A., Hoeve, M. A., ten Dam, M., van der Burg, M., van Dongen, J., . . . Lankester, A. (2004). The same IkappaBalpha mutation in two related individuals leads to completely different clinical syndromes. *Journal of Experimental Medicine, 200*(5), 559-568. doi:10.1084/jem.20040773

Jasny, B. R., Zahn, L. M., & Marshall, E. (2009). Complex systems and networks. Connections. Introduction. *Science, 325*(5939), 405. doi:10.1126/science.325_405

Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature, 411*(6833), 41-42. doi:Doi 10.1038/35075138

Jiang, H., & Chess, L. (2009). How the Immune System Achieves Self-Nonself Discrimination During Adaptive Immunity. *Advances in Immunology, Vol 102, 102*, 95-133. doi:10.1016/S0065-2776(09)01202-4

Joyce, A. R., & Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology, 7*(3), 198-210. doi:10.1038/nrm1857

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research, 40*(Database issue), D109-114. doi:10.1093/nar/gkr988 [doi]

Karaca, E., Karakoc-Aydiner, E., Bayrak, O. F., Keles, S., Sevli, S., Barlan, I. B., . . . Ozen, M. (2013). Identification of a novel mutation in ZAP70 and prenatal diagnosis in a Turkish family with severe combined immunodeficiency disorder. *Gene, 512*(2), 189-193. doi:10.1016/j.gene.2012.10.062

Keerthikumar, S., Bhadra, S., Kandasamy, K., Raju, R., Ramachandra, Y. L., Bhattacharyya, C., . . . Pandey, A. (2009). Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA Research, 16*(6), 345-351. doi:10.1093/dnares/dsp019

Kim, Y., Sidney, J., Buus, S., Sette, A., Nielsen, M., & Peters, B. (2014). Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics, 15*, 241. doi:10.1186/1471-2105-15-241

Klamt, S., Saez-Rodriguez, J., & Gilles, E. D. (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology, 1*, 2. doi:10.1186/1752-0509-1-2

Klamt, S., Saez-Rodriguez, J., Lindquist, J. A., Simeoni, L., & Gilles, E. D. (2006). A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics, 7*, 56. doi:10.1186/1471-2105-7-56

Klingström, T., & Plewczynski, D. (2011). Protein-protein interaction and pathway databases, a graphical review. *Brief Bioinform, 12*(6), 702-713. doi:10.1093/bib/bbq064

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., . . . Ogasawara, N. (2003). Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences of the United States of America, 100*(8), 4678-4683. doi:10.1073/pnas.0730515100

Kohl, M., Wiese, S., & Warscheid, B. (2011). Cytoscape: Software for Visualization and Analysis of Biological Networks. *Methods in Molecular Biology, 696*, 291-303. doi:10.1007/978-1-60761-987-1_18

Krumsiek, J., Poelsterl, S., Wittmann, D. M., & Theis, F. J. (2010). Odefy - From discrete to continuous models. *BMC Bioinformatics, 11*, 233-233. doi:10.1186/1471-2105-11-233

Kung, C., Pingel, J. T., Heikinheimo, M., Klemola, T., Varkila, K., Yoo, L. I., . . . Thomas, M. L. (2000). Mutations in the tyrosine phosphatase CD45 gene in a child with severe combined immunodeficiency disease. *Nature Medicine, 6*(3), 343-345. doi:10.1038/73208

Kurosaki, T., Shinohara, H., & Baba, Y. (2010). B Cell Signaling and Fate Decision. *Annual Review of Immunology, Vol 28, 28*, 21-55. doi:10.1146/annurev.immunol.021908.132541

Kurtz, J. (2004). Memory in the innate and adaptive immune systems. *Microbes and Infection, 6*(15), 1410-1417. doi:10.1016/j.micinf.2004.10.002

Le, D. H., & Kwon, Y. K. (2011). NetDS: a Cytoscape plugin to analyze the robustness of dynamics and feedforward/feedback loop structures of biological networks. *Bioinformatics, 27*(19), 2767-2768. doi:10.1093/bioinformatics/btr466

LeBien, T. W., & Tedder, T. F. (2008). B lymphocytes: how they develop and function. *Blood, 112*(5), 1570-1580. doi:10.1182/blood-2008-02-078071

Linka, R. M., Risse, S. L., Bienemann, K., Werner, M., Linka, Y., Krux, F., . . . Borkhardt, A. (2012). Loss-of-function mutations within the IL-2 inducible kinase ITK in patients with EBV-associated lymphoproliferative diseases. *Leukemia, 26*(5), 963-971. doi:10.1038/leu.2011.371

Linnemann, C., van Buuren, M. M., Bies, L., Verdegaal, E. M. E., Schotte, R., Calis, J. J. A., . . . Schumacher, T. N. M. (2015). High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4(+) T cells in human melanoma. *Nature Medicine, 21*(1), 81-85. doi:10.1038/nm.3773

Liu, Y., Shepherd, E. G., & Nelin, L. D. (2007). MAPK phosphatases--regulating the immune response. *Nature Reviews: Immunology, 7*(3), 202-212. doi:10.1038/nri2035

Lucas, C. L., Kuehn, H. S., Zhao, F., Niemela, J. E., Deenick, E. K., Palendira, U., . . . Uzel, G. (2014). Dominant-activating germline mutations in the gene encoding the PI(3)K catalytic subunit p110delta result in T cell senescence and human immunodeficiency. *Nature Immunology, 15*(1), 88-97. doi:10.1038/ni.2771

Luce, R. D., & Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika, 14*(2), 95-116.

Malissen, B., Aguado, E., & Malissen, M. (2005). Role of the LAT adaptor in T-cell development and Th2 differentiation. *Advances in Immunology, 87*, 1-25. doi:10.1016/S0065-2776(05)87001-4

Markowetz, F., & Spang, R. (2007). Inferring cellular networks--a review. *BMC Bioinformatics, 8 Suppl 6*, S5. doi:10.1186/1471-2105-8-S6-S5

Matsushita, H., Sato, Y., Karasaki, T., Nakagawa, T., Kume, H., Ogawa, S., . . . Kakimi, K. (2016). Neoantigen Load, Antigen Presentation Machinery, and Immune Signatures Determine Prognosis in Clear Cell Renal Cell Carcinoma. *Cancer Immunology Research, 4*(5), 463-471. doi:10.1158/2326-6066.Cir-15-0225

McDonald, D. R., Mooster, J. L., Reddy, M., Bawle, E., Secord, E., & Geha, R. S. (2007). Heterozygous N-terminal deletion of IkappaBalpha results in functional nuclear factor kappaB haploinsufficiency, ectodermal dysplasia, and immune deficiency. *Journal of Allergy and Clinical Immunology, 120*(4), 900-907. doi:10.1016/j.jaci.2007.08.035

Medzhitov, R., & Janeway, C. A., Jr. (2002). Decoding the patterns of self and nonself by the innate immune system. *Science, 296*(5566), 298-300. doi:10.1126/science.1068883

Mendoza, L. (2006). A network model for the control of the differentiation process in Th cells. *BioSystems, 84*(2), 101-114. doi:10.1016/j.biosystems.2005.10.004

Mendoza, L., & Pardo, F. (2010). A robust model to describe the differentiation of T-helper cells. *Theory Biosci, 129*(4), 283-293. doi:10.1007/s12064-010-0112-x

Mendoza, L., & Xenarios, I. (2006). A method for the generation of standardized qualitative dynamical systems of regulatory networks. *Theoretical Biology & Medical Modelling, 3*, 13. doi:10.1186/1742-4682-3-13

Mishra, N., Schreiber, R., Stanton, I., & Tarjan, R. E. (2007). Clustering social networks. *Algorithms and Models for the Web-Graph, 4863*, 56-+.

Mitchell, S., Vargas, J., & Hoffmann, A. (2016). Signaling via the NFkappaB system. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 8*(3), 227-241. doi:10.1002/wsbm.1331

Morgan, N. V., Goddard, S., Cardno, T. S., McDonald, D., Rahman, F., Barge, D., . . . Maher, E. R. (2011). Mutation in the TCRalpha subunit constant gene (TRAC) leads to a human immunodeficiency disorder characterized by a lack of TCRalphabeta+ T cells. *Journal of Clinical Investigation, 121*(2), 695-702. doi:10.1172/JCI41931

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H. H., . . . Sette, A. (2006). A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nature Biotechnology, 24*(7), 817-819. doi:DOI 10.1038/nbt1215

Murphy, K., Janeway, C. A., Travers, P., Walport, M., Mowat, A., & Weaver, C. T. (2012). *Janeway's immunobiology*. London; New York: Garland Science. Taylor & Francis Group.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E, 74*(3). doi:10.1103/PhysRevE.74.036104

Newman, M. E. J. (2010). *Networks: An Introduction.*: Oxford University Press 2010.

Oh-hora, M., & Rao, A. (2008). Calcium signaling in lymphocytes. *Current Opinion in Immunology, 20*(3), 250-258. doi:10.1016/j.coi.2008.04.004

Ortutay, C., & Vihinen, M. (2009a). Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Research, 37*(2), 622-628. doi:10.1093/nar/gkn982

Ortutay, C., & Vihinen, M. (2009b). Immunome knowledge base (IKB): an integrated service for immunome research. *BMC Immunology, 10*, 3. doi:10.1186/1471-2172-10-3

Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., . . . Wu, C. J. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature, 547*(7662), 217-221. doi:10.1038/nature22991

Pal, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics, 37*(12), 1372-1375. doi:10.1038/ng1686

Palacios, E. H., & Weiss, A. (2004). Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene, 23*(48), 7990-8000. doi:10.1038/sj.onc.1208074

Pancer, Z., & Cooper, M. D. (2006). The evolution of adaptive immunity. *Annual Review of Immunology, 24*, 497-518. doi:10.1146/annurev.immunol.24.021605.090542

Panda, S., & Ding, J. L. (2015). Natural Antibodies Bridge Innate and Adaptive Immunity. *Journal of Immunology, 194*(1), 13-20. doi:10.4049/jimmunol.1400844

Pannicke, U., Baumann, B., Fuchs, S., Henneke, P., Rensing-Ehl, A., Rizzi, M., . . . Schwarz, K. (2013). Deficiency of innate and acquired immunity caused by an IKBKB mutation. *New England Journal of Medicine, 369*(26), 2504-2514.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., . . . Brazma, A. (2011). ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research, 39*(Database issue), D1002-1004. doi:10.1093/nar/gkq1040

Picard, C., Al-Herz, W., Bousfiha, A., Casanova, J. L., Chatila, T., Conley, M. E., . . . Gaspar, H. B. (2015). Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. *Journal of Clinical Immunology*. doi:10.1007/s10875-015-0201-1

Picard, C., Dogniaux, S., Chemin, K., Maciorowski, Z., Lim, A., Mazerolles, F., . . . Hivroz, C. (2009). Hypomorphic mutation of ZAP70 in human results in a late onset immunodeficiency and no autoimmunity. *European Journal of Immunology, 39*(7), 1966-1976. doi:10.1002/eji.200939385

Piirilä, H., Väliaho, J., & Vihinen, M. (2006). Immunodeficiency mutation databases (IDbases). *Human Mutation, 27*(12), 1200-1208. doi:10.1002/humu.20405 [doi]

Pritchard, A. L., Burel, J. G., Neller, M. A., Hayward, N. K., Lopez, J. A., Fatho, M., . . . Schmidt, C. W. (2015). Exome Sequencing to Predict Neoantigens in Melanoma. *Cancer Immunol Res, 3*(9), 992-998. doi:10.1158/2326-6066.CIR-15-0088

R-Core-Team. (2016). R: A Language and Environment for Statistical Computing. Retrieved from http://www.R-project.org

Radicchi, F., Ramasco, J. J., & Fortunato, S. (2011). Information filtering in complex weighted networks. *Physical Review. E: Statistical, Nonlinear, and Soft Matter Physics, 83*(4 Pt 2), 046101. doi:10.1103/PhysRevE.83.046101

Razick, S., Magklaras, G., & Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics, 9*, 405. doi:10.1186/1471-2105-9-405

Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., & Marsh, S. G. (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research, 43*(Database issue), D423-431. doi:10.1093/nar/gku1161

Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., . . . Schraven, B. (2007). A logical model provides insights into T cell receptor signaling. *PLoS Computational Biology, 3*(8), e163. doi:10.1371/journal.pcbi.0030163

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B. P., Simon, P., Lower, M., . . . Tureci, O. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature, 547*(7662), 222-226. doi:10.1038/nature23003

Samarghitean, C., Ortutay, C., & Vihinen, M. (2009). Systematic classification of primary immunodeficiencies based on clinical, pathological, and laboratory parameters. *Journal of Immunology, 183*(11), 7569-7575. doi:10.4049/jimmunol.0901837

Samarghitean, C., Väliaho, J., & Vihinen, M. (2007). IDR knowledge base for primary immunodeficiencies. *Immunome Research, 3*, 6. doi:10.1186/1745-7580-3-6

Santoni, D., Pedicini, M., & Castiglione, F. (2008). Implementation of a regulatory gene network to simulate the TH1/2 differentiation in an agent-based model of hypersensitivity reactions. *Bioinformatics, 24*(11), 1374-1380. doi:10.1093/bioinformatics/btn135

Sawabe, T., Horiuchi, T., Nakamura, M., Tsukamoto, H., Nakahara, K., Harashima, S. I., . . . Nakano, S. (2001). Defect of lck in a patient with common variable immunodeficiency. *International Journal of Molecular Medicine, 7*(6), 609-614.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., . . . Ye, J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 40*(Database issue), D13-25. doi:10.1093/nar/gkr1184

Schroeder, M. L., Triggs-Raine, B., & Zelinski, T. (2016). Genotyping an immunodeficiency causing c.1624-11G>A ZAP70 mutation in Canadian Mennonites. *BMC Medical Genetics, 17*(1), 50. doi:10.1186/s12881-016-0312-4

Schumacher, T. N., & Hacohen, N. (2016). Neoantigens encoded in the cancer genome. *Current Opinion in Immunology, 41*, 98-103. doi:10.1016/j.coi.2016.07.005

Schumacher, T. N., & Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science, 348*(6230), 69-74. doi:10.1126/science.aaa4971

Sementchenko, V. I., & Watson, D. K. (2000). Ets target genes: past, present and future. *Oncogene, 19*(55), 6533-6548. doi:10.1038/sj.onc.1204034

Serrano, M. A., Boguna, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America, 106*(16), 6483-6488. doi:10.1073/pnas.0808904106

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., . . . Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America, 102*(6), 1974-1979. doi:DOI 10.1073/pnas.0409522102

Shields, R. (2012). Cultural Topology: The Seven Bridges of Konigsburg, 1736. *Theory Culture & Society, 29*(4-5), 43-57. doi:10.1177/0263276412451161

Smith-Garvin, J. E., Koretzky, G. A., & Jordan, M. S. (2009). T cell activation. *Annual Review of Immunology, 27*, 591-619. doi:10.1146/annurev.immunol.021908.132706

Song, C. M., Havlin, S., & Makse, H. A. (2005). Self-similarity of complex networks. *Nature, 433*(7024), 392-395. doi:10.1038/nature03248

Stepensky, P., Keller, B., Buchta, M., Kienzler, A.-K., Elpeleg, O., Somech, R., . . . Warnatz, K. (2013). Deficiency of caspase recruitment domain family, member 11 (CARD11), causes profound combined immunodeficiency in human subjects. *Journal of Allergy and Clinical Immunology, 131*(2), 477-+. doi:10.1016/j.jaci.2012.11.050

Stepensky, P., Weintraub, M., Yanir, A., Revel-Vilk, S., Krux, F., Huck, K., . . . Resnick, I. B. (2011). IL-2-inducible T-cell kinase deficiency: clinical presentation and therapeutic approach. *Haematologica, 96*(3), 472-476. doi:10.3324/haematol.2010.033910

Stojmirovic, A., & Yu, Y. K. (2011). ppiTrim: constructing non-redundant and up-to-date interactomes. *Database: The Journal of Biological Databases and Curation, 2011*, bar036. doi:10.1093/database/bar036

Supek, F., Bosnjak, M., Skunca, N., & Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One, 6*(7), e21800. doi:10.1371/journal.pone.0021800

Swain, S. L., Bradley, L. M., Croft, M., Tonkonogy, S., Atkins, G., Weinberg, A. D., . . . Huston, G. (1991). Helper T-Cell Subsets - Phenotype, Function and the Role of Lymphokines in Regulating Their Development. *Immunological Reviews, 123*, 115-144. doi:DOI 10.1111/j.1600-065X.1991.tb00608.x

Taminau, J., Meganck, S., Lazar, C., Steenhoff, D., Coletta, A., Molter, C., . . . Nowe, A. (2012). Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. *BMC Bioinformatics, 13*, 335. doi:10.1186/1471-2105-13-335

Tchilian, E. Z., Wallace, D. L., Wells, R. S., Flower, D. R., Morgan, G., & Beverley, P. C. (2001). A deletion in the gene encoding the CD45 antigen in a patient with SCID. *Journal of Immunology, 166*(2), 1308-1313.

Thomas, M. L., & Brown, E. J. (1999). Positive and negative regulation of Src-family membrane kinases by CD45. *Immunology Today, 20*(9), 406-411. doi:S0167-5699(99)01506-6

Thome, M. (2004). CARMA1, BCL-10 and MALT1 in lymphocyte development and activation. *Nature Reviews: Immunology, 4*(5), 348-359. doi:10.1038/nri1352

Torres, J. M., Martinez-Barricarte, R., Garcia-Gomez, S., Mazariegos, M. S., Itan, Y., Boisson, B., . . . Perez de Diego, R. (2014). Inherited BCL10 deficiency impairs hematopoietic and nonhematopoietic immunity. *Journal of Clinical Investigation, 124*(12), 5239-5248. doi:10.1172/JCI77493

Tran, E., Robbins, P. F., & Rosenberg, S. A. (2017). 'Final common pathway' of human cancer immunotherapy: targeting random somatic mutations. *Nature Immunology, 18*(3), 255-262. doi:10.1038/ni.3682

Trolle, T., Metushi, I. G., Greenbaum, J. A., Kim, Y., Sidney, J., Lund, O., . . . Nielsen, M. (2015). Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics, 31*(13), 2174-2181. doi:10.1093/bioinformatics/btv123

Tumminello, M., Aste, T., Di Matteo, T., & Mantegna, R. N. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America, 102*(30), 10421-10426. doi:10.1073/pnas.0500298102

Turvey, S. E., Durandy, A., Fischer, A., Fung, S. Y., Geha, R. S., Gewies, A., . . . Warnatz, K. (2014). The CARD11-BCL10-MALT1 (CBM) signalosome complex: Stepping into the limelight of human primary immunodeficiency. *Journal of Allergy and Clinical Immunology, 134*(2), 276-284. doi:10.1016/j.jaci.2014.06.015

van der Burg, M., & Gennery, A. R. (2011). The expanding clinical and immunological spectrum of severe combined immunodeficiency. *European Journal of Pediatrics, 170*(5), 561-571. doi:10.1007/s00431-011-1452-3

Verdegaal, E. M., de Miranda, N. F., Visser, M., Harryvan, T., van Buuren, M. M., Andersen, R. S., . . . van der Burg, S. H. (2016). Neoantigen landscape dynamics during human melanoma-T cell interactions. *Nature, 536*(7614), 91-95. doi:10.1038/nature18945

Vihinen, M. (2015). Immunodeficiency, Primary: Affecting the Adaptive Immune System. In *eLS*. Chichester: John Wiley & Sons Ltd, Chichester.

Vormehr, M., Diken, M., Boegel, S., Kreiter, S., Tureci, O., & Sahin, U. (2016). Mutanome directed cancer immunotherapy. *Current Opinion in Immunology, 39*, 14-22. doi:10.1016/j.coi.2015.12.001

Walpole, J., Papin, J. A., & Peirce, S. M. (2013). Multiscale Computational Models of Complex Biological Systems. *Annual Review of Biomedical Engineering, Vol 15, 15*, 137-154. doi:10.1146/annurev-bioeng-071811-150104

Wang, D., Matsumoto, R., You, Y., Che, T., Lin, X. Y., Gaffen, S. L., & Lin, X. (2004). CD3/CD28 costimulation-induced NF-kappaB activation is mediated by recruitment of protein kinase C-theta, Bcl10, and IkappaB kinase beta to the immunological synapse through CARMA1. *Molecular and Cellular Biology, 24*(1), 164-171.

Wang, R. S., & Albert, R. (2011). Elementary signaling modes predict the essentiality of signal transduction network components. *BMC Systems Biology, 5*, 44. doi:10.1186/1752-0509-5-44

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442. doi:Doi 10.1038/30918

Willmann, K. L., Klaver, S., Dogu, F., Santos-Valente, E., Garncarz, W., Bilic, I., . . . Boztug, K. (2014). Biallelic loss-of-function mutation in NIK causes a primary immunodeficiency with multifaceted aberrant lymphoid immunity. *Nat Commun, 5*, 5360. doi:10.1038/ncomms6360

Wilson, C. L., & Miller, C. J. (2005). Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics, 21*(18), 3683-3685. doi:10.1093/bioinformatics/bti605

Wittmann, D. M., Krumsiek, J., Saez-Rodriguez, J., Lauffenburger, D. A., Klamt, S., & Theis, F. J. (2009). Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. *BMC Systems Biology, 3*, 98. doi:10.1186/1752-0509-3-98

Yasuda, T., Sanjo, H., Pages, G., Kawano, Y., Karasuyama, H., Pouyssegur, J., . . . Kurosaki, T. (2008). Erk kinases link pre-B cell receptor signaling to transcriptional events required for early B cell expansion. *Immunity, 28*(4), 499-508. doi:10.1016/j.immuni.2008.02.015

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics, 26*(7), 976-978. doi:10.1093/bioinformatics/btq064

Zanudo, J. G. T., & Albert, R. (2013). An effective network reduction approach to find the dynamical repertoire of discrete dynamic networks. *Chaos, 23*(2). doi:10.1063/1.4809777

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research, 33*(Web Server issue), W741-748. doi:10.1093/nar/gki475

Teku et al. BMC Systems Biology 2014, 8:17
http://www.biomedcentral.com/1752-0509/8/17

BMC Systems Biology

RESEARCH ARTICLE Open Access

Identification of core T cell network based on immunome interactome

Gabriel N Teku[1†], Csaba Ortutay[2,3†] and Mauno Vihinen[1,2,3*]

Abstract

Background: Data-driven studies on the dynamics of reconstructed protein-protein interaction (PPI) networks facilitate investigation and identification of proteins important for particular processes or diseases and reduces time and costs of experimental verification. Modeling the dynamics of very large PPI networks is computationally costly.

Results: To circumvent this problem, we created a link-weighted human immunome interactome and performed filtering. We reconstructed the immunome interactome and weighed the links using jackknife gene expression correlation of integrated, time course gene expression data. Statistical significance of the links was computed using the Global Statistical Significance (GloSS) filtering algorithm. P-values from GloSS were computed for the integrated, time course gene expression data. We filtered the immunome interactome to identify core components of the T cell PPI network (TPPIN). The interconnectedness of the major pathways for T cell survival and response, including the T cell receptor, MAPK and JAK-STAT pathways, are maintained in the TPPIN network. The obtained TPPIN network is supported both by Gene Ontology term enrichment analysis along with study of essential genes enrichment.

Conclusions: By integrating gene expression data to the immunome interactome and using a weighted network filtering method, we identified the T cell PPI immune response network. This network reveals the most central and crucial network in T cells. The approach is general and applicable to any dataset that contains sufficient information.

Keywords: Protein-protein interaction, Network, Filtering, T cell, TPPIN, Signaling, PPI

Background

Cellular interactomes often consist of large numbers of proteins with even larger numbers of connections between them. Typically in protein-protein interaction (PPI) network nodes represent proteins and the links represent relationships between them. This network representation enables the study and visualization of the reconstructed cellular systems.

Data-driven studies on the dynamics of reconstructed PPI networks facilitate investigation and identification of proteins important for a particular process and reduces time and costs of experimental verification [1,2]. Modeling the dynamics of very large PPI networks is computationally very costly. To circumvent this problem, one needs to identify relevant core components of networks without losing vital information. A PPI network constituting most of the relevant core of a cellular system is sufficient to study its dynamic properties [3].

Many methods have been developed to reduce complex directed and undirected networks to their core components. Some of the methods include topological centrality techniques [4], synthetic biology approaches of the minimal gene set of a cell [5,6], complex systems coarse-graining [7,8], and filtering approaches [9-11]. In the centrality methods, topological centrality of nodes is used to identify the non-redundant links and to delete the redundant ones [11]. Minimal gene set approaches aim to identify genes that are crucial for life sustenance and cannot be inactivated under specific optimal growth conditions. These approaches do not take into account interactions between essential gene products [5]. The coarse-graining approaches identify specific motifs in a

* Correspondence: mauno.vihinen@med.lu.se
†Equal contributors
[1]Department of Experimental Medical Science, Lund University, Lund, Sweden
[2]Institute of Biomedical Technology, University of Tampere, Tampere, Finland
Full list of author information is available at the end of the article

network, and collapse and replace them by a single node [8]. This process is repeated until there are no more motifs. The final network is less complex but does not consider the structural heterogeneity and broad weight distribution, i.e. the multi-scale nature, of cellular networks.

Network filtering approaches have also been used to reduce network complexity [10-13]. Those that preserve the inherent multiscale structure of natural complex networks have been shown to be better in revealing most of the important components of networks [11,13]. These approaches score the nodes or links, and enable the deletion of those that do not deviate significantly from a null model.

In this study, we identified the network of proteins relevant in T cells by filtering the immunome interactome using the result from Global Statistical Significance (GloSS) [13] algorithm and a constraint of connectivity of the T cell receptor (TCR) signaling pathway. We compiled genes for the major immune processes and used them to reconstruct the immunome interactome, i.e., all the PPIs of the immunome. We then integrated gene expression profiles for the corresponding genes across several experiments. Jackknife correlation for gene expression was then used to weigh links between the proteins encoded by the genes. To maintain the multiscale structure of the network during filtering, we used the GloSS algorithm. This algorithm utilizes a global null model of the link weight and the degree distribution of the network. It computes the statistical significance for each link. For the null model, GloSS assigns weights from the weight distribution of the network, independently and randomly, without changing its topology. We filtered the network by deleting links based on their p-values (computed by GloSS) in descending order. To determine the endpoint of the filtering, we imposed as a constraint, the existence of a single path between the components of the NF-κB and TCR complexes.

Because we investigated the global and aggregate characteristics of the system and integrated T cell gene expressions, we can assume that the filtered network contains most of the components central for T cell signaling [14]. This was supported by Gene Ontology (GO) and essential genes enrichment analysis.

Results

Protein-protein interaction network

We used altogether 1579 proteins for the network filtering (Additional file 1). Eight hundred and eighty five human immunome genes were obtained from the Immunome Knowledge Base (IKB) [15]. As IKB contains only the most essential immunome genes and does not necessarily contain full pathways, it was supplemented with proteins for key immune system pathways derived from the KEGG Pathway database [16] (Table 1).

Table 1 KEGG pathways used to supplement IKB dataset

KEGG identifier	Name of KEGG pathway
path:hsa04010	MAPK signaling pathway
path:hsa04062	Chemokine signaling pathway
path:hsa04514	Cell adhesion molecules
path:hsa04612	Antigen processing and presentation
path:hsa04620	Toll-like receptor signaling pathway
path:hsa04621	NOD-like receptor signaling pathway
path:hsa04622	RIG-1-like receptor signaling pathway
path:hsa04630	Jak-STAT signaling pathway
path:hsa04640	Hematopoietic cell lineage
path:hsa04650	Natural killer cell mediated cytotoxicity
path:hsa04660	T cell receptor signaling pathway
path:hsa04662	B cell receptor signaling pathway
path:hsa04664	FcεRI signaling pathway
path:hsa04666	FcγR-mediated phagocytosis
path:hsa04670	Leukocyte trans-endothelial migration
path:hsa04672	Intestinal immune network for IgA production
path:hsa04610	Complement and coagulation cascades
path:hsa04623	Cytosolic DNA-sensing pathway

The protein products of the genes that take part in these pathways were used to supplement the protein data from the IKB database. The combined protein data represent the immune response protein dataset.

The PPI network was reconstructed for the immunome proteins (see workflow in Figure 1). PPI data were retrieved from iRefIndex database (version 9.0) which compiles PPIs from the major repositories [17]. ppiTrim (version 1.2.1) was used for general filtering according to Stojmirovic et al. [18]. Only experimentally verified and binary PPIs were retained. Moreover, multiple binary PPIs encoded by the same gene pair were collapsed into a single PPI. Finally, binary interactions to proteins outside the immunome were eliminated. A total of 5603 PPIs between 1259 immunome proteins were available after these pre-processing steps (Additional files 2 and 3).

Gene expression correlation

T cell gene expression datasets were obtained from NCBI GEO [19] and EBI ArrayExpress [20] databases. Altogether 16 time series datasets (Additional file 4) containing 384 samples derived from 5 platforms fulfilled the set criteria. After pre-processing, batch effect analysis was performed. Further, exploratory Principal Component Analysis (PCA) was done to examine the effect and performance of the batch effect analysis (Figure 2). The samples cluster according to experiment and platform before batch effect analysis. However, after batch effect correction, samples performed on all three platforms overlap with each other. The batch effect-corrected expression data were integrated

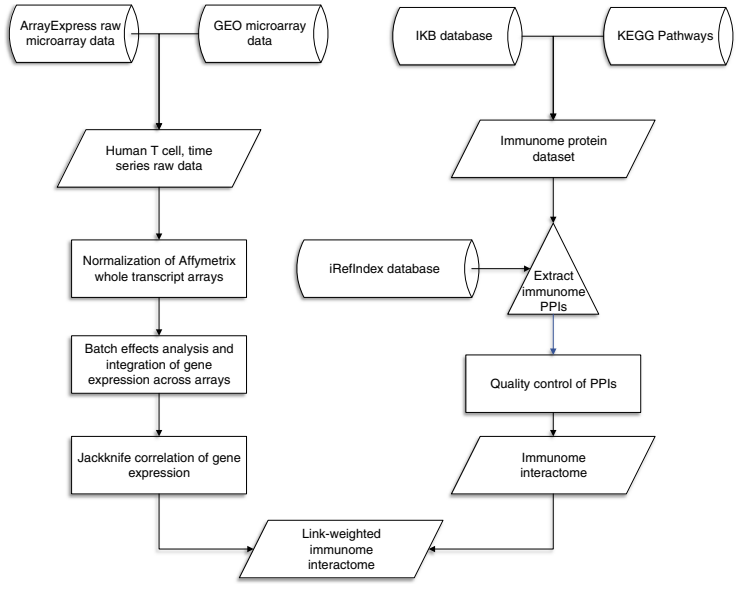


Figure 1 Workflow for the reconstruction of the immunome interactome. The general steps taken to reconstruct the immunome interactome are shown. The cylinders represent data repositories from which data was retrieved. Parallelograms represent data, either retrieved from databases or obtained by analyses. Rectangles represent performed analyses. T cell microarray experiments available in ArrayExpress and the GEO databases were retrieved. These experiments included at least 3 samples. The selected experiments were normalized using the R/Bioconductor libraries. Batch effect analysis was done and all experiments were merged or integrated. Jackknife Pearson correlation coefficient was calculated for the integrated dataset. Immunome proteins were retrieved from the Immunome knowledge Base (IKB) and the KEGG pathways databases. Major immune response pathways from the KEGG were used to supplement the IKB immunome proteins. Immunome interactome was obtained by retrieving PPIs for the immunome protein dataset from the iRefIndex database. To reduce noisy PPIs we used the ppiTrim method and further filtered its output of redundant and non-immune response PPIs. The Jackknife correlation coefficients were used as link weights.

or merged together. Of the genes encoding the 1259 immunome proteins, 1149 were expressed in at least 80% of the samples in the merged dataset and were thus included in the analysis.

Next, the mean of the jackknife Pearson product-moment correlation coefficient was calculated for the pre-processed and merged expression values for all gene pair combinations. In total, 1140 genes representing 5164 gene pairs encoding interacting proteins in the immunome interactome were used for further analysis.

The distribution of the integrated jackknife correlation values is shown in Figure 3. The maximum gene expression correlation is 0.88, between *ITGA2B* (integrin α-IIb or *CD41*) and *ITGB3* (integrin β-3 or *CD61*). The encoded proteins form an integrin receptor complex [21] and are thus co-expressed. Their functions include cell adhesion, cell-cell interaction, receptor for several molecules and platelet activation [21]. The minimum correlation of -0.62 was observed between *LCK*, coding for lymphocyte-specific protein tyrosine kinase, and *PAK2*, p21 protein (Cdc42/Rac)-activated kinase 2. LCK is an important signaling protein in many cellular processes, especially in T cell receptor (TCR) activation and T cell development [22]. PAK2 is a member of the PAK proteins (a family of serine/threonine kinases) targeted by small GTP proteins, CDC42 and RAC1 [23,24]. They take part in several signaling pathways, including the TCR signaling network. Albeit association of increased PAK2 activity in cells that overexpress Src kinases, PAK2 and LCK have not been shown to directly interact with each other [25]. The mean of the correlation values for all gene pairs is 0.09 and most of the correlation coefficients lie between -0.5 and 0.5.

T cell-specific PPI network

We reconstructed the immunome PPI network as a weighted and undirected graph. The nodes, links, and link weights of the graph represent, respectively, the immunome protein coding genes, the PPIs and the absolute value of the mean jackknife expression correlation between the connected immunome protein coding genes.

The topology and weight distribution of naturally occurring complex weighted networks are heterogeneous

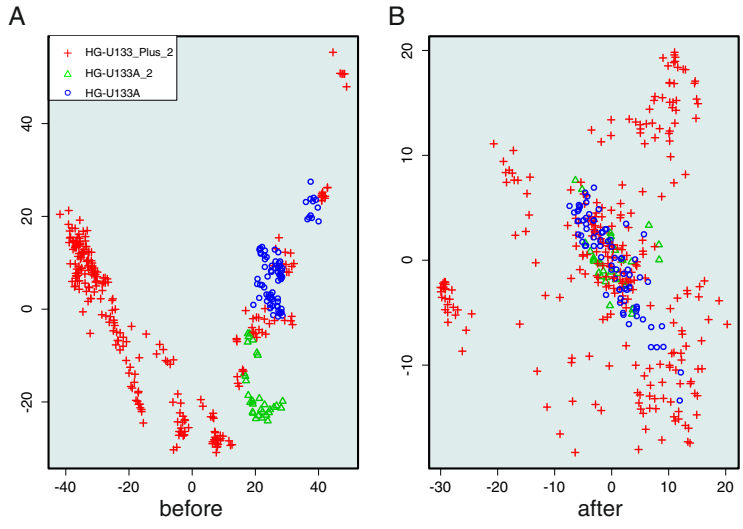


Figure 2 PCA analysis of normalized gene expression data before and after batch effect analysis. A. PCA before batch effect analysis. Experiments from the different platforms cluster together. **B**. PCA after batch effect analysis. Results for experiments from different platforms overlap. The platforms are indicated by symbols.

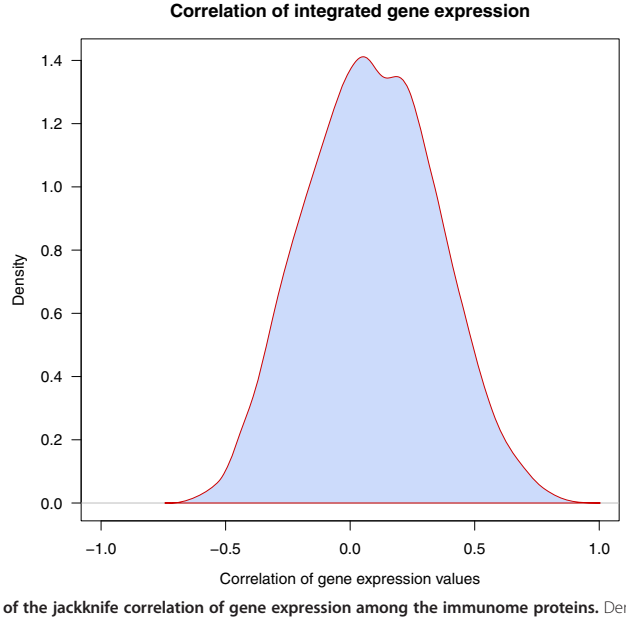


Figure 3 The distribution of the jackknife correlation of gene expression among the immunome proteins. Density plot for the distribution of gene expression correlation of 1140 immune response-related genes used in this study. Gene expression values were normalized, and integrated across experiments after batch effect analysis. The correlation coefficient for each gene pair was derived by the jackknife Pearson correlation coefficient across all the integrated microarray samples. A major part of the gene expression correlation values is between -0.5 and 0.5.

and tightly connected. This makes the identification of the relevant structure that maintains the multiscale nature of the network nontrivial. Thus, we used the GloSS algorithm [13] to compute a p-value, for each link. GloSS identifies the relevant backbone of a weighted graph while retaining the multiscale coupling of its weight distribution and topological characteristics. It uses a global null model that describes both the structure of the network and its weight distribution. The p-values computed by GloSS were used to filter the network by deleting links based on their p-values, in descending order. We monitored the filtering process to make sure that the central networks between TCR, and NF-κB and NFAT signaling pathways remained intact. These pathways have been shown to be crucial for T cell signaling [26,27] and therefore cannot be disconnected without destroying essential cellular processes.

We followed changes of structural and biological features in the PPI network during the filtering process with network parameters. The diameter of the network represents the longest minimum distance between the nodes. We used as measures the changes in diameter, the relative size of the largest connected component and the average size of the isolated components [28]. These network topology scores show how connectivity, integrity and robustness of the network are changed when links are removed during the filtering process (Figure 4). All the panels in Figure 4 indicate that at the cutoff point most of the remaining network's connectivity and integrity is still maintained. We call the remaining network the T cell PPI Network, TPPIN (Figure 5). TPPIN consists of 288 nodes, 227 links in 73 connected components (Table 2).

Correlation distribution before and after filtering

Threshold algorithms filter a network by removing edges whose weights are below an arbitrary cutoff. Such a network loses its multiscale and, thus, its core structure. We probed the distribution of the gene expression correlation coefficient to establish whether the multiscale structure of the immunome interactome is retained in the filtered T cell PPI network (Figure 6). The filtering process succeeds in maintaining not just the links with large weights but also links with lower weights. Thus, the filtering process maintains the multi-scale structure of the network and retains edges that are crucial for the T cell PPI network.

Effect of noise on the filtering procedure

To test the sensitivity of our filtering procedure to noise we introduced randomness to the immunome interactome, before performing filtering, by randomizing fractions of the link weights while preserving the topology of the network. We refer to these networks as the Link Weight-Randomized Networks (LWRNs). Nine such networks were created based on the fraction of weights randomized. Thirty iterations were conducted for each LWRN. Each iteration consists of choosing randomly a fraction of links, reassigning their weights randomly, conducting the filtering procedure, and calculating network topology statistics. The topology features calculated for each iteration include node degree, average path length, betweenness centrality of both the nodes and the links, clustering coefficient of the network, and the intersection between the TPPIN and the LWRN. These measures indicate the local and global connectivity of a network. We retained the average of the above quantities.

Figure 7 shows the similarity or dissimilarity between TPPIN and LWRNs. Figure 7 A-E, shows that as more of the link weights are randomized, the topology of the LWRNs diverges significantly from TPPIN. Moreover, as Figure 7 F shows, there is very little overlap of links between the LWRNs and TPPIN.

Gene Ontology over-representation and semantic similarity analysis

GO term over-representation analysis was performed for the TPPIN proteins and shows that, at level two details, most of the biological process terms are relevant for T cell function (Table 3 and Additional file 5). For example, the term *positive regulation of lymphocyte activation pathway* (GO:0051251, p-value = 9.74×10^{-7}), *regulation of immune response* (GO:0050776, p-value = 1.11×10^{-6}), and *intracellular protein kinase cascade* (GO:0007243, p-value = 3.40×10^{-6}) terms are among the most significantly enriched after adjusting for multiple comparisons. In addition to significant immune response-related terms, there are also those for general cellular processes.

To better investigate the similarity or difference between the immunome interactome and the TPPIN network, we explored semantic similarity of the networks using the GOSemSim package available from R/Bioconductor. The semantic similarity ranges between 0 and 1. The similarity between the immunome interactome and TPPIN proteins in the biological process and molecular function terms were very high, i.e., 0.91 and 0.92, respectively, indicating that the TPPIN is very representative of the immunome interactome.

Essential genes over-representation analysis

Essential genes are indispensable to the survival of a cell or organism. To account for how essential the genes are, we performed an over-representation analysis to identify the proportion of the essential TPPIN genes. We conducted a hypergeometric test on the human orthologs of the mouse lethality genes from the Mouse Genome Informatics resource [29]. The results show a highly

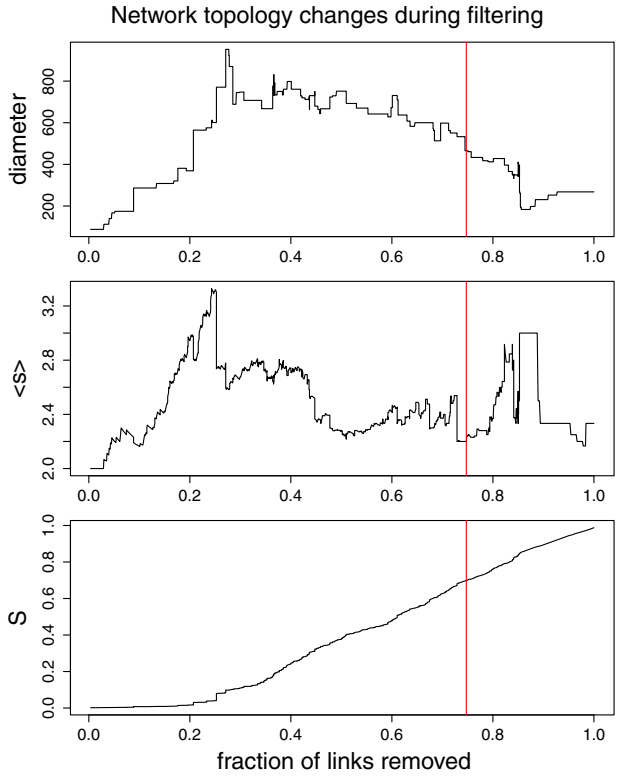


Figure 4 Network topology changes during GloSS-, NFAT- and NF-κB-assisted filtering. The immune response PPI network topological changes during the filtering process. Network measures were used to investigate the immunome interactome during filtering. The x-axis in each panel is the fraction of nodes removed during filtering. On the y-axis, the top panel shows changes in the network diameter, the middle panel changes in the average size of the isolated components excluding the largest or giant component (<s>). The bottom panel shows changes in the relative size of the largest or giant component (S). The relative size of the largest component is the number of nodes in the largest component divided by the number of nodes in the whole network. That is, $n_{rel} = n/N$, where n_{rel} is the relative size of the largest component, n is the number of nodes in the largest component and N is the number of nodes in the whole network). Each of the network measures were plotted against the fraction of links removed during filtering. The vertical line shows the point at which the paths between the TCR complex and the NF-κB and NFAT downstream components are broken. This also represents the point at which the filtering process stops. This indicates that the connectivity and robustness of the filtered network at this endpoint is maintained. Thus the connectivity and robustness inherent in the immunome interactome is maintained in the TPPIN.

significant enrichment of essential genes in the TPPIN (p-value = 1.37×10^{-10}, Table 4 and Figure 5).

Interconnection of T cell-specific pathways

The TPPIN proteins were mapped onto the TCR, JAK-STAT and MAPK signaling pathways that are central for T cell functions [30] (Figure 8). Albeit containing just a third of the proteins in the initial network, the TPPIN includes almost all the main components for the remaining pathways. Except for CD3γ and CD3δ, all the CD3 proteins of the TCR complex are present in the TPPIN. Further, most proteins important for early T cell activation, NFAT, AP1, NF-κB, T cell co-inhibitory and co-stimulatory signal transduction are present. Overall, most of the proteins in the important pathways for T cell signaling are present in the TPPIN. This indicates that the filtering procedure was able to, first of all, identify central pathways and, secondly, to keep their connectivity. As a novel feature the TPPIN indicates the interconnection of the central pathways.

Discussion and conclusions

In this study, we identified the network of proteins relevant for T cells by filtering the multiscale immunome

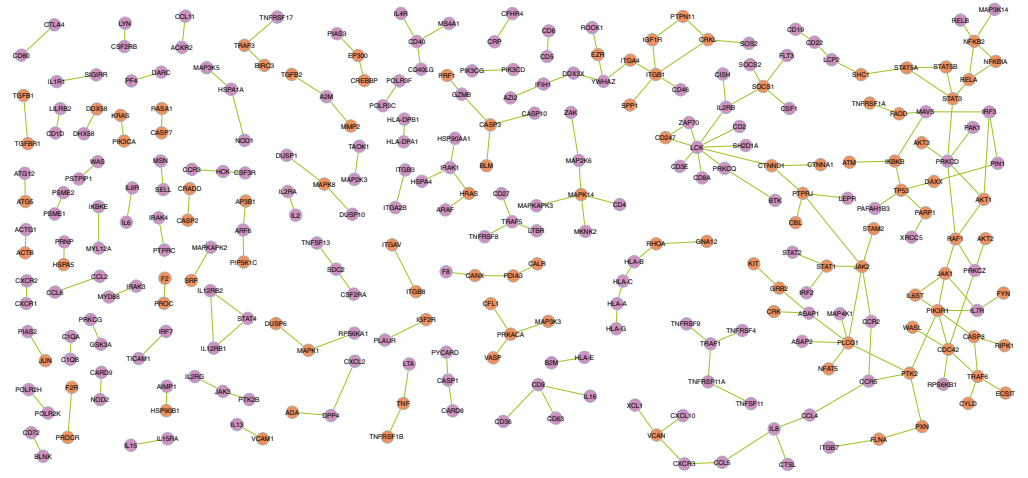


Figure 5 T cell PPI Network and essential genes enrichment. After removing non-immunome genes and immunome genes without the lethality annotations, we calculated the hypergeometric distribution and Fisher's exact test for significance. The figure demonstrates the enrichment of essential genes across almost every component of the TPPIN and for the whole network (p-value = 1.37×10^{-10}). The node labels are gene identifiers for the genes coding the proteins. The nodes colored red represent essential genes. The essential genes data is based on the human orthologs of the mouse lethality genes from the Mouse Genome Informatics database.

interactome using the GloSS filtering algorithm [13]. We compiled the genes for the major immune processes and reconstructed the immunome interactome. Then we integrated gene expression profiles across several gene expression experiments. The jackknife correlation for gene expression was used to weigh links between the proteins encoded by the genes. Next, we used the output from GloSS to filter the network. The filtered network contains most of the relevant T cell functional components and was designated TPPIN. This was confirmed by the overrepresentation analysis conducted with GO terms and essential genes.

Many important components of the TCR-dependent signaling pathways are present in the TPPIN. Except for CD3γ and CD3δ, other components of the TCR complex which are included in the microarrays used in this study, are present (TCR-α and -β are not present in the microarrays). The co-receptors CD4 and CD8 are both present, as well as, all the proteins that make up the immunological synapse. With the exception of LAT, GADS and ITK, most proteins that are central in the immediate TCR receptor-associated intracellular signaling after the formation of the immunological synapse and TCR activation are present in the TPPIN, including LCK, FYN, CD45, ZAP70, SLP-76 and PLC-γ.

After its activation, PLC-γ cleaves PIP2 into the second messenger IP3 and DAG [31,32]. This event sets off the activation of three important signaling pathways in T cells that end up with transcriptional activation of NFAT, NF-κB and AP-1 [30]. DAG activates PKC-θ, which in turn activates NF-κB [33]. IP3 activates CaN through the calcium signaling, and CaN subsequently activates NFAT [34]. DAG activates RasGRP [35,36], which in turn initiates the activation of the MAP kinase cascade [37], culminating in the activation of FOS [38]. Key proteins in the NF-κB pathway including PKC-θ, IKK-β and IκB [39] are present in the TPPIN. With the exception of RasGRP, MEK1/2 and ELK co-complexes, the other vital proteins in the MAP kinase signaling cascade [40] and the JAK-STAT pathway [41] are captured by the TPPIN. These results show how the TPPIN represents relevant T cell-related parts of the immunome interactome.

Table 2 General structure of the T cell PPI network

Number of nodes in connected component	Number of links	Number of components in the network
91	100	1
14	14	1
6	5	2
5	4	3
4	3	5
3	2	13
3	3	1

A component represents a set of nodes that are all connected to each other, either directly or indirectly. Components with two nodes are not included in the table.

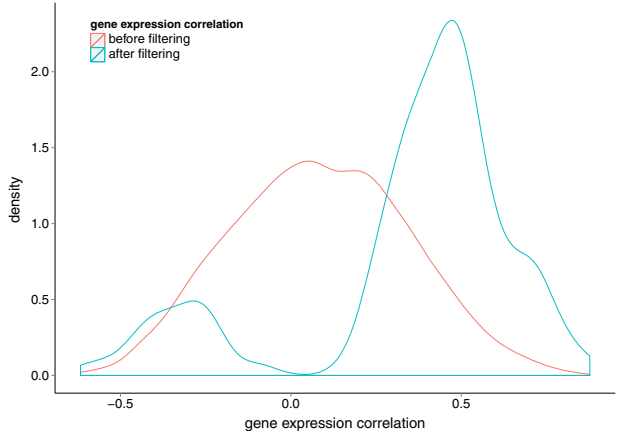


Figure 6 Distribution of correlation coefficient before and after filtering. Threshold algorithms filter a network by removing edges whose weights are below an arbitrary cutoff. Such a network would have lost its multiscale structure and thus its core structure. We probe the distribution of the gene expression correlation coefficient to establish whether the multiscale structure of the immunome interactome is retained in the filtered T cell PPI network. The red and blue curves represent, respectively, the distribution of gene expression correlation before and after filtering. The filtering approach preserves a broad distribution of link weights, i.e., most with large weights and some with small weights.

During the filtering step the central networks connecting the TCR complex to the NF-κB and NFAT signaling pathways were kept intact. Although the NFAT and NF-κB pathways are present in many different cell-types, they are central for T cell survival and functions. The connectivity of these components was used to determine the end point for the filtering process. The filtering was continued until there was a minimum number of links, i.e., one, between the TCR, and NF-κB and NFAT components.

GO term enrichment analysis confirms that several of the TPPIN proteins have important T cell functions. As an example of biological process term enrichment, the *positive regulation of lymphocyte activation pathway* (GO:0051251), *regulation of immune response* (GO:0050776), and *intracellular protein kinase cascade* (GO:0007243) terms are significantly enriched. To further probe the similarity between the immunome interactome and the TPPIN proteins we calculated their semantic similarity with respect to biological process and molecular function GO terms. The networks were semantically very similar in both types of GO terms. Because essential genes are indispensable for the survival of a cell, their enrichment in the cellular network would indicate that the network is crucial to the cell. Thus, we investigated the enrichment of essential genes in the TPPIN. The analysis showed a highly significant enrichment of essential genes in the TPPIN. These independent lines of evidence support the applicability of the network filtering routine.

Due to the scarcity of time course microarray experiments with uniform design, gene expression datasets with different designs were used. Integrated analysis was carried out to identify and exclude biased datasets [42,43]. The normalization and batch effect analysis steps served to considerably minimize the effect of bias for correlation calculation from the experimental studies.

Global and aggregate cellular interactions are more plausible between proteins encoded by co-expressed genes than between gene products whose expression patterns are uncorrelated [14]. Since we investigated the global and aggregated characteristics of the immune response in T cells by integrating gene expression experiments conducted for T cell lines, the correlation coefficients represent the aggregate strength of the T cell-specific relationship between the genes and their interacting protein products [14,44].

To explore the changes in the network during the filtering process we investigated changes in the diameter, relative size of the largest component and the average size of the connected components of the network. These network measures have been shown to indicate the connectivity status of a network and its robustness against link removal or loss [28,45]. The changes in network statistics during the filtering process showed that TPPIN maintains most of the integrity and connectivity of the immunome interactome.

Certain aspects of T cell function have been previously modeled [46-49]. Most of these studies are related to gene regulatory networks and modeling of small signaling networks involving transcription factors and their targets, selected to include genes or proteins well-known

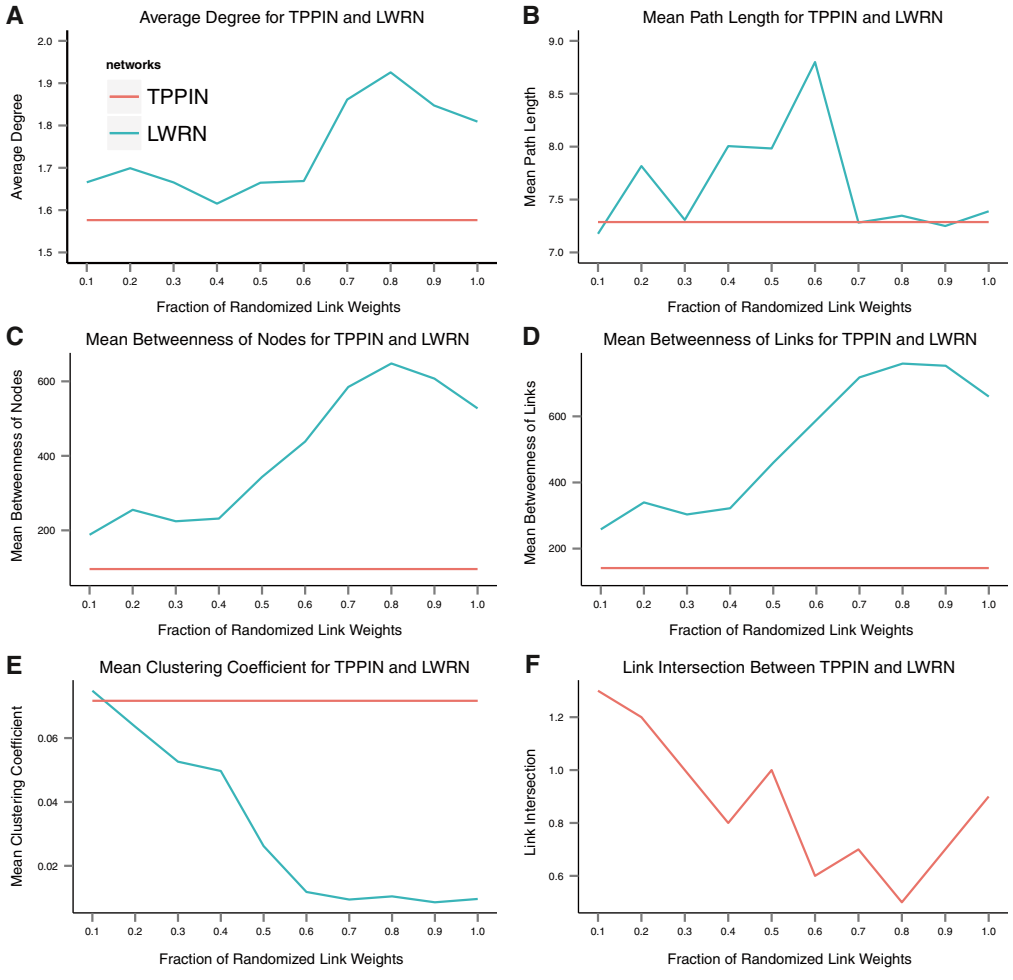


Figure 7 Comparison of T cell PPI network and link weight-randomized networks. Robustness analysis on the filtering procedure was performed by randomizing the immunome interactome link weights to introduce noise. Thirty iterations were conducted so that a fraction of links are selected randomly and their weights randomly reassigned for each iteration. The Link Weight-Randomized Network (LWRN) was filtered and node degree, average path length, betweenness centrality of both the nodes and the links, clustering coefficient of the network, and the intersection between TPPIN and the LWRNs, were calculated. The blue line curves are for the LWRNs and red lines for the TPPIN. **A**. Change in average degree, **B**. change in mean path length, **C**. change in mean betweenness of nodes, **D**. change in mean betweenness of links, **E**. change in mean clustering coefficient, **F**. Link intersection between TPPIN and LWRNs.

in the modeled system. In these studies, the typical number of genes or proteins is in a few tens, whereas we started with the entire immunome interactome of 1149 proteins and 5164 links, and ended up with a core network that contains 288 proteins and 227 links. The number of nodes and links in the TPPIN makes it amenable to tailored cellular systems modeling and experimental studies. Our approach is unsupervised and does not utilize any preconceptions, yet, it reveals the central proteins and their networks.

The filtering process carried out in this study has some potential limitations. It needs several time course expression datasets for the cell-type or tissue of interest and each experiment should consist of at least 3 samples. A set of proteins is needed to track the connectivity of the vital pathways and a stop criterion when key pathways are

Table 3 GO *biological process* term enrichment for TPPIN

GO ID	Term	Number of significant vs. annotated genes	Expected number of genes	Raw vs. adjusted P value
GO:0051251	positive regulation of lymphocyte activation	128/61	32.51	$5.40 \times 10^{-09}/9.74 \times 10^{-07}$
GO:0043067	regulation of programmed cell death	289/114	73.41	$4.77 \times 10^{-10}/4.60 \times 10^{-07}$
GO:0050776	regulation of immune response	313/118	79.51	$6.79 \times 10^{-09}/1.11 \times 10^{-06}$
GO:0048523	negative regulation of cellular process	401/142	101.86	$1.05 \times 10^{-08}/1.62 \times 10^{-06}$
GO:0050867	positive regulation of cell activation	144/64	36.58	$7.04 \times 10^{-08}/5.59 \times 10^{-06}$
GO:0048584	positive regulation of response to stimulus	401/140	101.86	$5.10 \times 10^{-08}/4.40 \times 10^{-06}$
GO:0042981	regulation of apoptotic process	285/113	72.39	$4.04 \times 10^{-10}/4.60 \times 10^{-07}$
GO:0007243	intracellular protein kinase cascade	330/121	83.82	$3.13 \times 10^{-08}/3.40 \times 10^{-06}$
GO:0019221	Cytokine mediated signaling pathway	163/73	41.40	$3.85 \times 10^{-09}/8.07 \times 10^{-07}$
GO:0006468	protein phosphorylation	313/116	79.51	$3.63 \times 10^{-08}/3.51 \times 10^{-06}$

The "universe" is the immunome protein data and the enrichment is for the filtered immunome interactome, the T cell PPI network (TPPIN).

broken. However, these limitations are not of great practical importance in the present era of high throughput studies.

The reported filtering routine can capture the core cell-type-specific PPI network for any cell-type from time series gene expression datasets, and is not limited to well-known systems. The approach opens ways for modeling protein interaction networks of cellular systems, even when pathways are not previously well characterized.

Methods

Protein-protein interaction network reconstruction

Human immunome proteins were obtained from the IKB [15] and supplemented with key immune system pathways from the KEGG pathways database [16].

Experimentally verified and consolidated PPI data for the human immunome proteins was retrieved from the iRefIndex database version 9.0 [17]. First, the ppiTrim version 1.2.1 [18] was used to filter the iRefIndex dataset. This algorithm maps protein interactants to NCBI gene identifiers and removes undesired raw interactions, deflates potentially expanded complexes, and reconciles annotation labels from the different PPI databases. Second, non-experimentally verified, non-human, complex and self-self PPIs were omitted. Third, we collapsed multiple binary PPIs whose interactants are products of the same genes. Finally, we eliminated PPIs for which both interactants were not immunome proteins (Figure 1). The igraph library [50] in the R statistical programming environment [51] was used to reconstruct and analyze the PPI network. Visualizations were done using Cytoscape version 2.8 [52].

Gene expression data

We retrieved microarray time course datasets for human T cell-lines from GEO [19] and ArrayExpress [20] databases. Each experiment had to contain at least three samples and at least one for time zero for baseline data. GEO datasets that consisted of samples from multiple platforms were split into multiple experiments, so that each experiment consisted of samples for the same microarray platform. To reduce bias during gene expression integration across experiments we included only experiments performed on Affymetrix whole transcript array platform U133A, U133A 2.0, U133B, U133 plus 2.0 and U95A arrays.

Pre-processing of gene expression data

R and Bioconductor libraries were used for data preprocessing [51,53]. The raw data for each gene expression dataset was retrieved. Pre-processing consisted of quality control using boxplots, arrayPLM and simpleaffy routines. For each experiment, samples were normalized using default parameters of the Robust Multi-Array algorithm [54] implemented in the affy library [55]. To convert probe sets to gene expressions, we used the mean of the probe sets to represent the corresponding gene's expression using the platform-dependent libraries in the Bioconductor project [56]. Gene expressions for

Table 4 Essential genes overrepresentation

	Number of genes	Number of genes annotated in MGI	Number of lethality genes annotated in MGI[a]	Expected number of lethality genes in MGI	P-value for hypergeometric test
Immunome interactome	1140	949	312		
TPPIN	288	256	105	59	1.37×10^{-10}

The T cell PPI network is the resulting network after filtering the immunome interactome. MGI[a] is the Mouse Genome Informatics database.

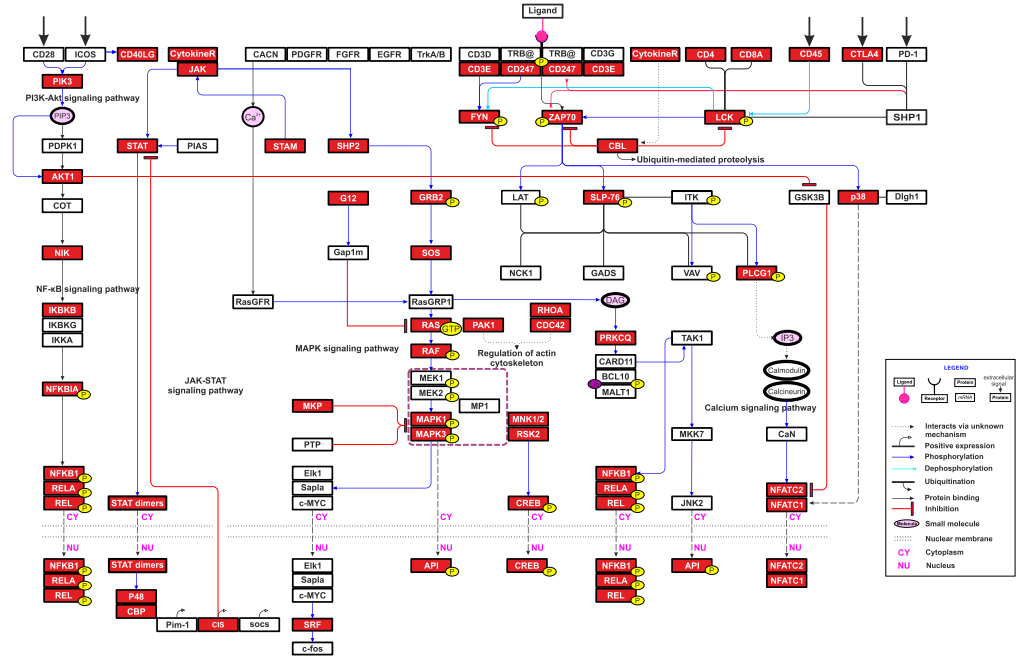


Figure 8 T cell PPI Network mapped to TCR, with main JAK-STAT and MAPK signaling pathways. The figure shows the TCR and the main JAK-STAT and MAPK signaling networks that are connected to the TCR-dependent T cell response. The pathway information is adapted from the KEGG Pathways resource. The TPPIN proteins are represented by red-colored boxes with protein names in white text. The signaling network demonstrates the effectiveness of the filtering approach as many of the central proteins in T cell response are left intact after filtering.

non-protein coding genes in the immunome protein dataset were removed.

The gene expression datasets were merged and batch effects were analyzed. We also performed PCA analysis before and after batch effect analysis to examine its effect and performance on the normalized datasets. The batch effects and PCA analysis were performed using the ComBat [43] and plotMDS algorithms implemented in the inSilicoMerging library [42] in Bioconductor.

Gene expression correlation

The mean of the jackknife Pearson correlation coefficient of the merged and pre-processed expression values for all gene pair combinations was calculated using the bootstrap library implemented in R. These correlation values were converted to absolute values and used as link weights for the immunome interactome.

Protein network filtering

We reconstructed the immunome PPI network as a weighted and undirected graph using the igraph package in R. The nodes, links, and link weights of the graph represent, respectively, the immunome protein coding genes, the PPIs and the average jackknife gene expression correlation between the immunome protein coding genes.

Network filtering was achieved with the GloSS algorithm [13], which identifies the relevant backbone of a weighted graph while retaining its weight distribution and structure. It uses a global null model to calculate the significance of the links by maintaining the topology of the network while assigning link weights randomly, from the observed weight distribution. The link weights (jackknife correlation coefficients) were multiplied by 100 to allow the p-values to be computed by GloSS. The computed link p-values by GloSS were used to filter the network by removing links in decreasing order of p-value. We monitored the filtering process to make sure that at least a path or connectivity remained between the TCR complex and NF-κB signaling pathways. The steps below represent the filtering procedure:

Step 1: Using GloSS, determine p-value for each edge of the network
Step 2: Select the link with the largest p-value
Step 3: Remove the link from the network

Step 4: Check for presence of connectivity between the NF-κB components and the TCR complex
Step 4.1: If connectivity exists discard the link and go to step 2.
Step 4.2: If connectivity does not exist, return the link to the network and stop.

This procedure was performed for both the NF-κB and the NFAT signaling pathways. Network diameter is the maximum of the shortest paths between the nodes of the network. A connected component is the region of a network in which there is a path connecting all node pairs. We followed changes in the network diameter, the relative size of the largest connected component and the average size of the isolated components [28]. The relative size of the largest component is the number of nodes in the largest component divided by the number of nodes in the whole network. That is, $n_{rel} = n/N$, where, n_{rel} is the relative size of the largest component, n is the number of nodes in the largest component and N is the number of nodes in the whole network. These measures were plotted against the fraction of filtered nodes. The ratio,

$$\frac{\text{number of deleted nodes}}{\text{number of nodes in the network}},$$

represents the fraction of the filtered nodes. The igraph package was used to calculate the network scores [50].

Robustness of the T cell PPI network

Link weight-randomized networks were created by randomizing the weights of a fraction of links, keeping the topology unchanged. The following fractions of links were used to create each of the link weight-randomized networks: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. Thirty iterations were performed on each link weight-randomized network. For each iteration, a fraction of links were randomly selected, their weights randomly reassigned, the filtering procedure performed and network topology statistics calculated. Node degree, average path length, betweenness centrality of both the nodes and the links, clustering coefficient of the network, and the intersection between the TPPIN network and the link weight-randomized networks, were calculated. After the iterations for each link weight-randomized network, the average of each of the network topology statistics was retained.

Gene Ontology term enrichment, over-representation and semantic similarity analysis

The interconnected proteins in the TPPIN were subjected to GO [57] term enrichment analysis. The GO terms for the proteins in the immunome interactome were used as the background. Fisher's exact test of the hypergeometric distribution was calculated and correction for multiple comparisons was performed using the Benjamini-Hochberg procedure [58]. The enrichment analysis was performed with Webgestalt [59]. Semantic similarity between the immunome interactome and the TPPIN was calculated using the clusterSim routine of the GOSemSim library [60] (version 1.18.0) available in R/Bioconductor.

Analysis of essential genes

We retrieved the human orthologs of the mouse lethality genes from the Mouse Genome Informatics database [29]. A gene was included in the set of lethality genes with the following criteria: phenotype contains the word "lethality", the type of lethality annotation contains neither "partial" nor "wean". After removing non-immunome genes and those without the above-mentioned lethality annotations, we calculated the hypergeometric distribution and Fisher's exact test for significance. Essential genes were retrieved using the biomaRt package in R [61] and visualization of the TPPIN with essential genes was done using Cytoscape 2.8.3.

Pathway gene mapping

The TPPIN genes were mapped to the KEGG pathways using the KEGG pathway mapper tool [16].

Additional files

Additional file 1: Protein data from the Immunome Knowledge Base and the immune response pathways from KEGG. This file contains the Entrez-gene identifiers of the genes encoding the immune response proteins from the IKB database and the KEGG immune response pathways listed in Table 1 of the main document. This dataset represents the immunome protein dataset and was used to generate the immunome interactome from PPIs in the iRefIndex database.

Additional file 2: Immunome interactome network figure. The figure represents the immunome interactome constructed from the immunome protein list of Additional file 1. The figure shows the complex nature of the network and thus cannot be studied by intuition alone. To reduce the complexity of the network the filtering procedure, reported in this study, was performed.

Additional file 3: Immunome interactome table. This is a table of the PPIs of the immune response proteins of Additional file 1. They were reconstructed from the iRefIndex which is a compendium of PPI data from major PPI databases. Additional filtering was carried out such that only experimentally verified, human, binary PPIs were obtained (see methods). The identifiers are entrez gene identifiers of the genes that code for the immune response genes.

Additional file 4: A summary of the gene expression datasets. This consists of a summary of all microarray datasets that were used in this study. The datasets were retrieved from NCBI's GEO and EBI's ArrayExpress databases. The dataset with asterisk (*) contains 3 experiments conducted on 3 different platforms. The 3 experiments were separated into separate data sets throughout the pre-processing. After pre-processing only samples from the experiment conducted on Affymetrix Human Genome U133A Array were merged with data sets from other experiments.

Additional file 5: Full Gene Ontology analysis results table. This contains details of the GO term enrichment analysis performed by the Webgestalt web resource. The background of the GO analysis is the immune response proteins. The null hypothesis significance test is the hypergeometric test and the p-values were corrected using the Benjamini–Hochberg procedure.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
GNT contributed towards data acquisition, analysis and interpretation; drafting and writing the manuscript. CO and MV contributed towards conception and design of this work; analysis and interpretation; drafting and writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Marko Pesu for valuable discussions.

Author details
[1]Department of Experimental Medical Science, Lund University, Lund, Sweden. [2]Institute of Biomedical Technology, University of Tampere, Tampere, Finland. [3]BioMediTech, University of Tampere, Tampere, Finland.

Received: 5 July 2013 Accepted: 5 February 2014
Published: 15 February 2014

References
1. Csermely P, Korcsmaros T, Kiss HJ, London G, Nussinov R: **Structure and dynamics of molecular networks: A novel paradigm of drug discovery: a comprehensive review.** *Pharmacol Ther* 2013, **138**(3):333–408.
2. Karlebach G, Shamir R: **Modelling and analysis of gene regulatory networks.** *Nat Rev Mol Cell Biol* 2008, **9**(10):770–780.
3. Kim JR, Kim J, Kwon YK, Lee HY, Heslop-Harrison P, Cho KH: **Reduction of complex signaling networks to a representative kernel.** *Sci Signal* 2011, **4:**175. ra35.
4. Newman ME: **Finding community structure in networks using the eigenvectors of matrices.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, **74**(3 Pt 2):036104.
5. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Le Coq D, Masson A, Mauel C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JF, Sekiguchi J, Sekowska A, Seror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M, Yamamoto Y, Yamane K, Yata K, Yoshida K, Yoshikawa H, Zuber U, Ogasawara N: **Essential *Bacillus subtilis* genes.** *Proc Natl Acad Sci U S A* 2003, **100**(8):4678–4683.
6. Commichau FM, Pietack N, Stulke J: **Essential genes in *Bacillus subtilis*: a re-evaluation after ten years.** *Mol Biosyst* 2013, **9**(6):1068–1075.
7. Song C, Havlin S, Makse HA: **Self-similarity of complex networks.** *Nature* 2005, **433**(7024):392–395.
8. Itzkovitz S, Levitt R, Kashtan N, Milo R, Itzkovitz M, Alon U: **Coarse-graining and self-dissimilarity of complex networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **71**(1 Pt 2):016127.
9. Santoni D, Pedicini M, Castiglione F: **Implementation of a regulatory gene network to simulate the TH1/2 differentiation in an agent-based model of hypersensitivity reactions.** *Bioinformatics* 2008, **24**(11):1374–1380.
10. Serrano MA, Boguna M, Vespignani A: **Extracting the multiscale backbone of complex weighted networks.** *Proc Natl Acad Sci U S A* 2009, **106**(16):6483–6488.
11. Grady D, Thiemann C, Brockmann D: **Robust classification of salient links in complex networks.** *Nat Commun* 2012, **3:**864.
12. Tumminello M, Aste T, Di Matteo T, Mantegna RN: **A tool for filtering information in complex systems.** *Proc Natl Acad Sci U S A* 2005, **102**(30):10421–10426.
13. Radicchi F, Ramasco JJ, Fortunato S: **Information filtering in complex weighted networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2011, **83**(4 Pt 2):046101.
14. Klebanov LB, Yakovlev AY: **A nitty-gritty aspect of correlation and network inference from gene expression data.** *Biol Direct* 2008, **3:**35.
15. Ortutay C, Vihinen M: **Immunome knowledge base (IKB): an integrated service for immunome research.** *BMC Immunol* 2009, **10:**3.
16. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**(1):D109–D114.
17. Razick S, Magklaras G, Donaldson IM: **iRefIndex: a consolidated protein interaction database with provenance.** *BMC Bioinforma* 2008, **9:**405.
18. Stojmirovic A, Yu YK: **ppiTrim: constructing non-redundant and up-to-date interactomes.** In *Database (Oxford) 2011.* ; 2011. bar036.
19. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2012, **40**(1):D13–D25.
20. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A: **ArrayExpress update-an archive of microarray and high-throughput sequencing-based functional genomics experiments.** *Nucleic Acids Res* 2011, **39**(Database issue):D1002–D1004.
21. Wippler J, Kouns WC, Schlaeger EJ, Kuhn H, Hadvary P, Steiner B: **The integrin αIIb-β3, platelet glycoprotein IIb-IIIa, can form a functionally active heterodimer complex without the cysteine-rich repeats of the β3 subunit.** *J Biol Chem* 1994, **269**(12):8754–8761.
22. Nakayama T, Yamashita M: **The TCR-mediated signaling pathways that control the direction of helper T cell differentiation.** *Semin Immunol* 2010, **22**(5):303–309.
23. Takino J, Yamagishi S, Takeuchi M: **Cancer malignancy is enhanced by glyceraldehyde-derived advanced glycation end-products.** *J Oncol* 2010, **2010:**739852.
24. Olivieri KC, Mukerji J, Gabuzda D: **Nef-mediated enhancement of cellular activation and human immunodeficiency virus type 1 replication in primary T cells is dependent on association with p21-activated kinase 2.** *Retrovirology* 2011, **8:**64–4690. 8-64.
25. Karkkainen S, Hiipakka M, Wang JH, Kleino I, Vaha-Jaakkola M, Renkema GH, Liss M, Wagner R, Saksela K: **Identification of preferred protein interactions by phage-display of the human Src homology-3 proteome.** *EMBO Rep* 2006, **7**(2):186–191.
26. Voll RE, Jimi E, Phillips RJ, Barber DF, Rincon M, Hayday AC, Flavell RA, Ghosh S: **NF-κB activation by the pre-T cell receptor serves as a selective survival signal in T lymphocyte development.** *Immunity* 2000, **13**(5):677–689.
27. Macian F: **NFAT proteins: key regulators of T-cell development and function.** *Nat Rev Immunol* 2005, **5**(6):472–484.
28. Albert R, Jeong H, Barabasi AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406**(6794):378–382.
29. Cox A, Ackert-Bicknell C, Dumont BL, Ding Y, Bell JT, Brockmann GA, Wergedal JE, Bult C, Paigen B, Flint J, Tsaih SW, Churchill GA, Broman KW: **A new standard genetic map for the laboratory mouse.** *Genetics* 2009, **182**(4):1335–1344.
30. Smith-Garvin JE, Koretzky GA, Jordan MS: **T cell activation.** *Annu Rev Immunol* 2009, **27:**591–619.
31. Berg LJ, Finkelstein LD, Lucas JA, Schwartzberg PL: **Tec family kinases in T lymphocyte development and function.** *Annu Rev Immunol* 2005, **23:**549–600.
32. Carpenter G, Ji Q: **Phospholipase C-γ as a signal-transducing element.** *Exp Cell Res* 1999, **253**(1):15–24.
33. Schmitz ML, Bacher S, Dienz O: **NF-κB activation pathways induced by T cell costimulation.** *FASEB J* 2003, **17**(15):2187–2193.
34. Hogan PG, Chen L, Nardone J, Rao A: **Transcriptional regulation by calcium, calcineurin, and NFAT.** *Genes Dev* 2003, **17**(18):2205–2232.

35. Ebinu JO, Bottorff DA, Chan EY, Stang SL, Dunn RJ, Stone JC: **RasGRP, a Ras guanyl nucleotide- releasing protein with calcium- and diacylglycerol-binding motifs.** *Science* 1998, **280**(5366):1082–1086.
36. Tognon CE, Kirk HE, Passmore LA, Whitehead IP, Der CJ, Kay RJ: **Regulation of RasGRP via a phorbol ester-responsive C1 domain.** *Mol Cell Biol* 1998, **18**(12):6995–7008.
37. Thomas G: **MAP kinase by any other name smells just as sweet.** *Cell* 1992, **68**(1):3–6.
38. Karin M, Liu Z, Zandi E: **AP-1 function and regulation.** *Curr Opin Cell Biol* 1997, **9**(2):240–246.
39. Weil R, Israel A: **Deciphering the pathway from the TCR to NF-κB.** *Cell Death Differ* 2006, **13**(5):826–833.
40. Rincon M: **MAP-kinase signaling pathways in T cells.** *Curr Opin Immunol* 2001, **13**(3):339–345.
41. Shuai K, Liu B: **Regulation of JAK-STAT signaling in the immune system.** *Nat Rev Immunol* 2003, **3**(11):900–911.
42. Taminau J, Meganck S, Lazar C, Steenhoff D, Coletta A, Molter C, Duque R, de Schaetzen V, Weiss Solis DY, Bersini H, Nowe A: **Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages.** *BMC Bioinforma* 2012, **13**:335–2105. 13-335.
43. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**(1):118–127.
44. Guo Y, Xiao P, Lei S, Deng F, Xiao GG, Liu Y, Chen X, Li L, Wu S, Chen Y, Jiang H, Tan L, Xie J, Zhu X, Liang S, Deng H: **How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes.** *Acta Biochim Biophys Sin (Shanghai)* 2008, **40**(5):426–436.
45. Cohen R, Erez K, ben Avraham D, Havlin S: **Resilience of the internet to random breakdowns.** *Phys Rev Lett* 2000, **85**(21):4626–4628.
46. Rui-Sheng W, Reka A: **Elementary signaling modes predict the essentiality of signal transduction network components.** *BMC Syst Biol* 2011, **5**:44.
47. Mendoza L, Pardo F: **A robust model to describe the differentiation of T-helper cells.** *Theory Biosci* 2010, **129**(4):283–293.
48. Mendoza L: **A network model for the control of the differentiation process in Th cells.** *BioSystems* 2006, **84**(2):101–114.
49. Mendoza L, Xenarios I: **A method for the generation of standardized qualitative dynamical systems of regulatory networks.** *Theor Biol Med Model* 2006, **3**:13.
50. Csardi G, Nepusz T: **The igraph software package for complex network research.** *Inter J, Complex Systems* 2006, **1**(1):1695.
51. **R: A Language and Environment for Statistical Computing.** http://www.r-project.org/.
52. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27**(3):431–432.
53. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80.
54. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**(2):249–264.
55. Gautier L, Cope L, Bolstad BM, Irizarry RA: **Affy-analysis of affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**(3):307–315.
56. **Bioconductor task view: annotation data.** http://www.bioconductor.org/packages/release/BiocViews.html#___AffymetrixChip.
57. The Gene Ontology Consortium: **The Gene Ontology: enhancements for 2011.** *Nucleic Acids Res* 2012, **40**(D1):D559–D564.
58. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B Stat Methodol* 1995, **57**(1):289–300.
59. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W741–W748.
60. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.** *Bioinformatics* 2010, **26**(7):976–978.
61. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nature Protocols* 2009, **4**(8):1184–1191.

doi:10.1186/1752-0509-8-17
Cite this article as: Teku *et al.*: Identification of core T cell network based on immunome interactome. *BMC Systems Biology* 2014 **8**:17.

PLOS ONE

RESEARCH ARTICLE

Simulation of the dynamics of primary immunodeficiencies in CD4+ T-cells

Gabriel N. Teku, Mauno Vihinen*

Department of Experimental Medical Science, Lund University, Lund, Sweden

* mauno.vihinen@med.lu.se

OPEN ACCESS

Citation: Teku GN, Vihinen M (2017) Simulation of the dynamics of primary immunodeficiencies in CD4+ T-cells. PLoS ONE 12(4): e0176500. https://doi.org/10.1371/journal.pone.0176500

Editor: Sunil K Ahuja, South Texas Veterans Health Care System, UNITED STATES

Received: January 23, 2017

Accepted: April 11, 2017

Published: April 27, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Vetenskaprådet, (http://www.vr.se/) to GNT and Barncancerfonden, (https://www.barncancerfonden.se) to VM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Primary immunodeficiencies (PIDs) form a large and heterogeneous group of mainly rare disorders that affect the immune system. T-cell deficiencies account for about one-tenth of PIDs, most of them being monogenic. Apart from genetic and clinical information, lots of other data are available for PID proteins and genes, including functions and interactions. Thus, it is possible to perform systems biology studies on the effects of PIDs on T-cell physiology and response. To achieve this, we reconstructed a T-cell network model based on literature mining and TPPIN, a previously published core T-cell network, and performed semi-quantitative dynamic network simulations on both normal and T-cell PID failure modes. The results for several loss-of-function PID simulations correspond to results of previously reported molecular studies. The simulations for TCR PTPRC, LCK, ZAP70 and ITK indicate profound changes to numerous proteins in the network. Significant effects were observed also in the BCL10, CARD11, MALT1, NEMO, IKKB and MAP3K14 simulations. No major effects were observed for PIDs that are caused by constitutively active proteins. The T-cell model facilitates the understanding of the underlying dynamics of PID disease processes. The approach confirms previous knowledge about T-cell signaling network and indicates several new important proteins that may be of interest when developing novel diagnosis and therapies to treat immunological defects.

Introduction

The human immunome consists of the genes and proteins essential both for the innate and adaptive immunity. Interactions between these proteins are indispensable for immune responses [1]. Studies have been carried out to identify and characterize the essential immunome interactome, i.e. the totality of interactions in the immune system [1, 2]. Knowledge from these studies enables the investigation of the dynamic behavior of networks in both health and disease. The immunome interactome varies depending on the cell-type, timing and localization of expressed and active proteins.

CD4+ T-cells are crucial immune response white blood cells. They recognize and bind to antigens on antigen-presenting cells via the cell surface T-cell receptor (TCR) complex [3]. Antigen binding to the TCR triggers a sequence of signaling events that lead to the activation and nuclear transportation of specific transcription factors (TFs) [3]. In the nucleus, these TFs

transactivate genes that are required for T-cell responses. CD4+ T-cells are divided into subpopulations of T helper 1 (Th1), Th2, Th17, regulatory T (Treg) and follicular helper T (Tfh) cells [4]. Each cell type plays different roles in the immune response by virtue of their different master regulator TFs and signature cytokine expression [5].

Here, we investigated the qualitative dynamics of the naïve CD4+ T-cells in both health and in disease in primary immunodeficiencies. Protein interaction networks in T-cells and their role in various diseases have been investigated [6, 7]. Primary immunodeficiencies (PIDs) are intrinsic diseases of the immune system, and are typically rare with heterogeneous phenotypes. Currently about 300 PIDs are known. Disease-causing variants in PIDs have been collected into the IDbases [8] and other databases and are available for more than 150 PIDs. Differential diagnosis of PIDs can be difficult due to overlapping signs and symptoms. Several classification schemes have been made, including the frequently updated classification by the International Union of Immunological Societies (IUIS) expert committee for PIDs [9]. PIDs have also been classified with a network approach that clusters the diseases based on signs, symptoms and laboratory parameters [10]. The severity of PIDs ranges from mild to moderate, and severe to lethal. By integrating the diverse information sources, systems level studies of the underlying mechanisms on PIDs can be conducted.

In systems biology, the reconstruction of cellular networks and their simulations facilitate studies of diseases as perturbations (or alterations) to the networks [11, 12]. These approaches provide insight on the dynamics of biomolecular interactions that drive cellular processes and contribute towards deciphering biological processes in both health and disease. Disease-causing variations can affect protein-protein interaction (PPI) networks at the cellular or tissue level. Studies of quantitative dynamics of PPIs require kinetic parameters and reaction constants. A problem emerges as reaction constants for most of the reactions have not been determined. Further, these network calculations are very computer intensive. The number of parameters, even for a moderate size network is so large that calculations would be very costly and time-consuming. Another approach amenable to larger networks of few tens to hundreds of nodes is to use qualitative and semi-quantitative dynamic methods [13–15], which provide useful models for approximating systems.

In this study, we employed a semi-quantitative method, the normalized HillCube Boolean approach [16], to simulate the dynamics during the activation of naïve CD4+ T-cells. With these simulations, we investigated the mechanisms of perturbations of known PID-causing proteins and revealed novel putative PID-related factors. Semi-quantitative simulations with synchronous updates were performed, and *in silico* validated. The simulations qualitatively replicated PIDs due to variations in PID-related proteins which disrupt essential signal transduction pathways during T-cell development from pre- to mature CD4+ T-cells [12]. Further, several novel proteins affected by PIDs were identified.

Results

The naïve CD4+ T-cell activation network

We reconstructed the signal transduction network for naïve CD4+ T-cells by using the T cell PPI network, TPPIN [17] as a basis for formulating reaction equations. The nodes and links of the PPI network were used to mine the published literature for valid reaction equations on CD4+ T-cells. The TPPIN is a PPI network that contains 227 core signal transduction interactions derived from integrated, time series, gene expression data sets. This network does not include link directions and, in most cases, lacks cellular context. Thus, we mined manually the direction, interaction and cellular context information by literature survey. We included only signaling interactions that were TCR/CD28-dependent and CD4+ T-cell-specific, leaving 85

interactions, which were used for reconstructing the Boolean network model (S1 Table). The interactions were defined manually as Boolean equations using the sum-of-product (SOP) form. The SOP representation offers a convenient means to represent Boolean equations of a signaling network in a hypergraph [18]. Proteins, i.e. the nodes, represent the Boolean variables. The edges (hyperarcs) represent the interactions between proteins and are signed either activating (+) or inhibiting (-). Edges have a tail that begins from a start node and a head (or arrow), which points to an end node, indicating the direction of signal transduction. Multiple edges with the same end node were summed by an OR operator. The AND gate was used as a product operator for multiple incoming edges that together are required to activate or inhibit a protein. 19 input nodes did not have in-coming links (Fig 1).

We started by analyzing the structure of the network and the signaling paths between the initial events of the TCR-dependent activation and the late events that involve the activation of the major TFs that turn on the expression of response genes. The TCR complex, its co-receptor CD4, and the co-stimulatory receptor CD28, are involved in the initial events, while the TFs AP1, NFAT, and NF-κB control the late events of T-cell activation [3].

The TCR is activated when it binds to an antigen (signal 1) presented by an antigen presenting cell. Another signal (signal 2) through the co-activation receptor CD28 is needed to elicit

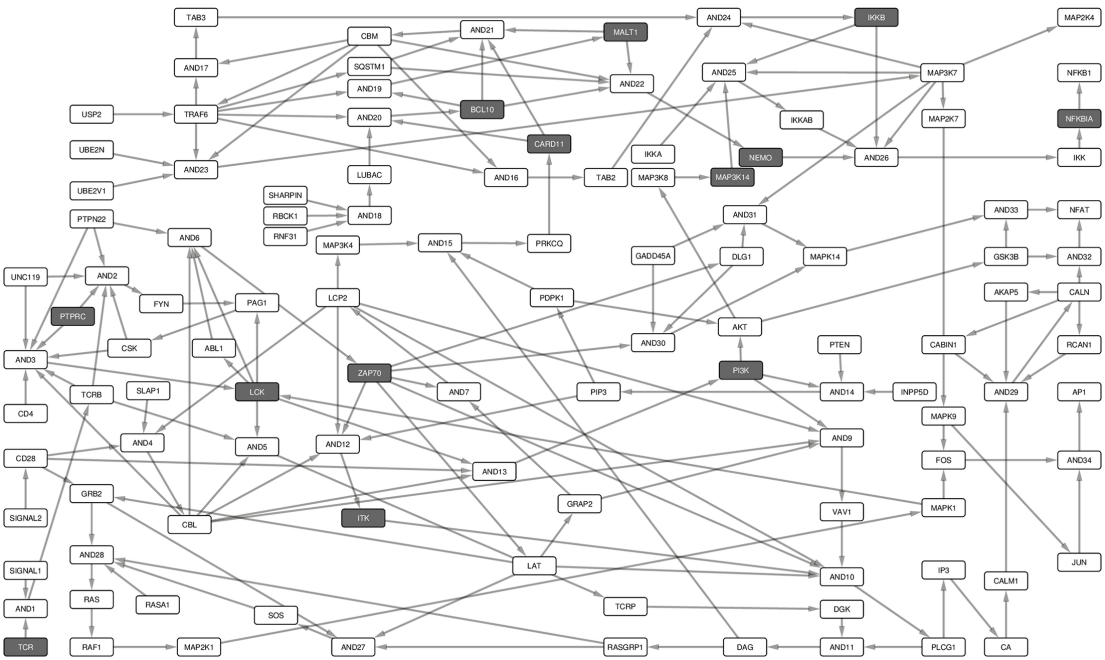


Fig 1. Naïve CD4+ T-cell activation Boolean network model. The network consists of 182 links and 118 nodes (including Boolean operators), 19 of which are input nodes, i.e., no link points to them (S1 Table). The Boolean network represents the naïve CD4+ T-cell activation events. The boxes represent non-PID (white) and PID proteins (gray). Spheres denote the AND gate. Activating links have a pointed head and solid line while inhibiting links have a blunt head and dashed line. Signal 1 represents peptide-MHC/TCR complex while Signal 2 represents co-receptor-ligand association, e.g. CD80-B7. Since the network focuses on TCR/CD28 signaling events, some events, e.g. for survival signaling that occur after antigen mediated T-cell activation and response through interleukin 2 (IL2), have not been fully considered.

https://doi.org/10.1371/journal.pone.0176500.g001

activation, survival and response [19]. The multiple paths from receptors to TFs guarantee a fail-safe and robust T-cell activation [20, 21]. It may also imply that the sensitivity of the level of activation is modulated by different routes [22]. On the other hand, signal transduction may be critical if only a single route exists from the receptors, through the network, to the TFs.

We identified signaling paths from signals 1 and 2 to the major response TFs NF-κB, AP1 and NFAT. For this purpose, we converted the Boolean network into an interaction network (Fig 2). Such a network captures the dependencies, interactions, and thus, the paths through which signals are transduced through the network. The interaction network consists of a connected component with 85 nodes interconnected by 146 links. To detect the part of the network with the most cross-talk between the signaling pathways, we identified the strongly connected component that consists of 25 nodes and 48 links (Fig 3).

To identify proteins essential for signal transduction from the receptors to the downstream actuators we analyzed the feedback loops (FBLs) in the network. Those proteins whose Boolean update equations are along most of the FBLs are considered essential. Input and output nodes were not included in the FBLs. We identified 419 such loops, of which the longest spans 20 nodes and the shortest 2 nodes (Fig 4). The median and mean length of the FBLs is 14 nodes long. Among the PID proteins, LCK was in 409 FBLs, ZAP70 in 380, CBM in 316, CARD11 in 312, BCL10 in 210, ITK in 120, PI3K in 110 and MALT1 in 106 FBLs. The other PID proteins, NEMO, IKKB, NFKBIA and MAP3K14, do not occur in any of the FBLs. PTPRC is an input node and is thus not included in any of the FBLs.

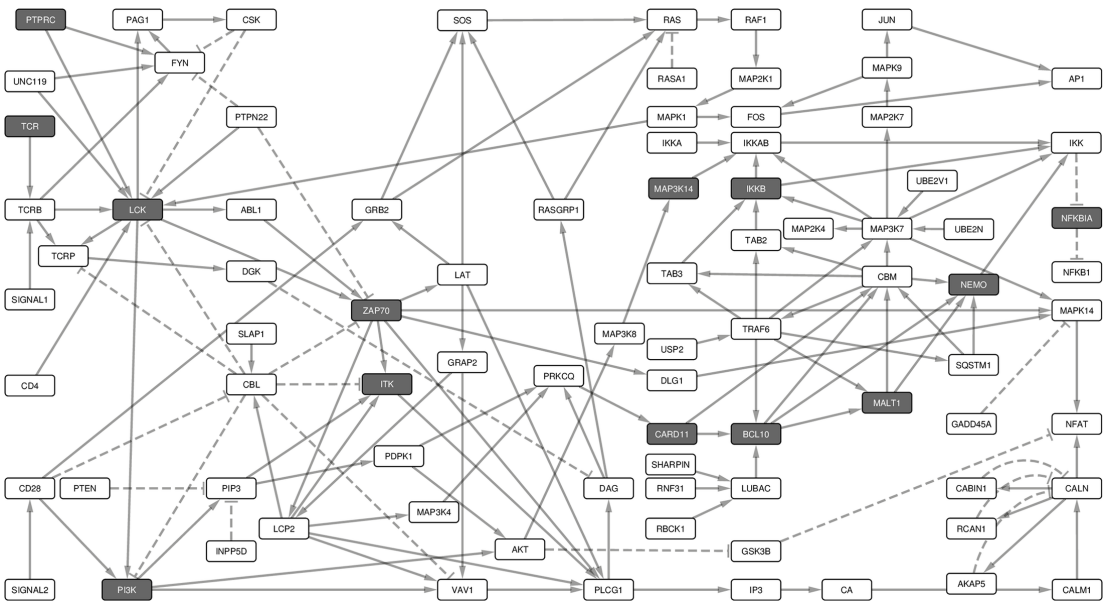


Fig 2. Boolean model transformed into its underlying interaction graph. The network consists of nodes and links derived from the Boolean network model without the AND operator. The interaction graph consists of 85 nodes and 146 links, and represents the underlying interaction network of the model. The nodes are as described in Fig 1. The network shows the paths through which signals from the receptors are channeled through the network to the TFs, which turn on the response genes.

https://doi.org/10.1371/journal.pone.0176500.g002

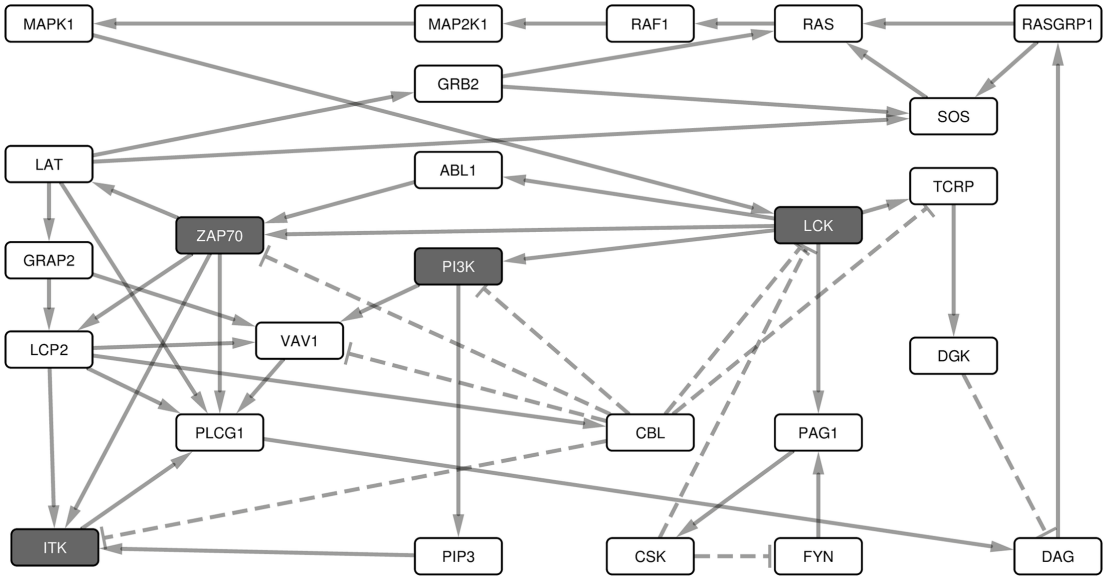


Fig 3. The strongly connected components of the interaction graph. The strongly connected component of the interaction graph consists of 25 nodes and 48 links. This subnet shows the interconnectedness and cross-talk of the early signals after the antigen-TCR ligation.

https://doi.org/10.1371/journal.pone.0176500.g003

Validation of reconstructed network and identification of the wild type attractor

After the engagement of the TCR complex and the co-activator CD28, a series of signal transduction cascades occur in naïve CD4+ T-cells [23] and are captured by the reconstructed network. The signaling cascades lead to response either via NF-κB, AP1 or NFAT [3]. The reconstructed network is cogent if the major TFs (here TFs NF-κB, AP1 and NFAT that together activate IL2) and the signaling components that lead to their activation are turned on.

To ensure that the reconstructed model reproduces CD4+ T-cell activation, we performed simulations by iteratively changing the initial states of the input nodes while making sure that the network represented the main signaling events. We used normalized HillCube dynamic simulations [16] with signals 1 and 2 turned on and validated the simulations *in silico*. This network model was used for the subsequent analyses. Additionally, we performed simulations by turning signal 1 on and signal 2 off, and vice versa. When either of the signals were turned off, only AP1 and NFAT, but not NF-κB, were activated.

The normalized HillCube simulations were run until the networks reached their attractor states. The model settled in a cyclic attractor or limit cycle after about 40 update rounds (arbitrary time units, Fig 5). The network subsequently continues in a cycle attractor after about every 20 updates. This attractor is in accordance with published experimental results [3, 24], which also is evident in the activation of the major downstream TFs (AP1, NF-κB, and NFAT) when signals 1 and 2 are turned on.

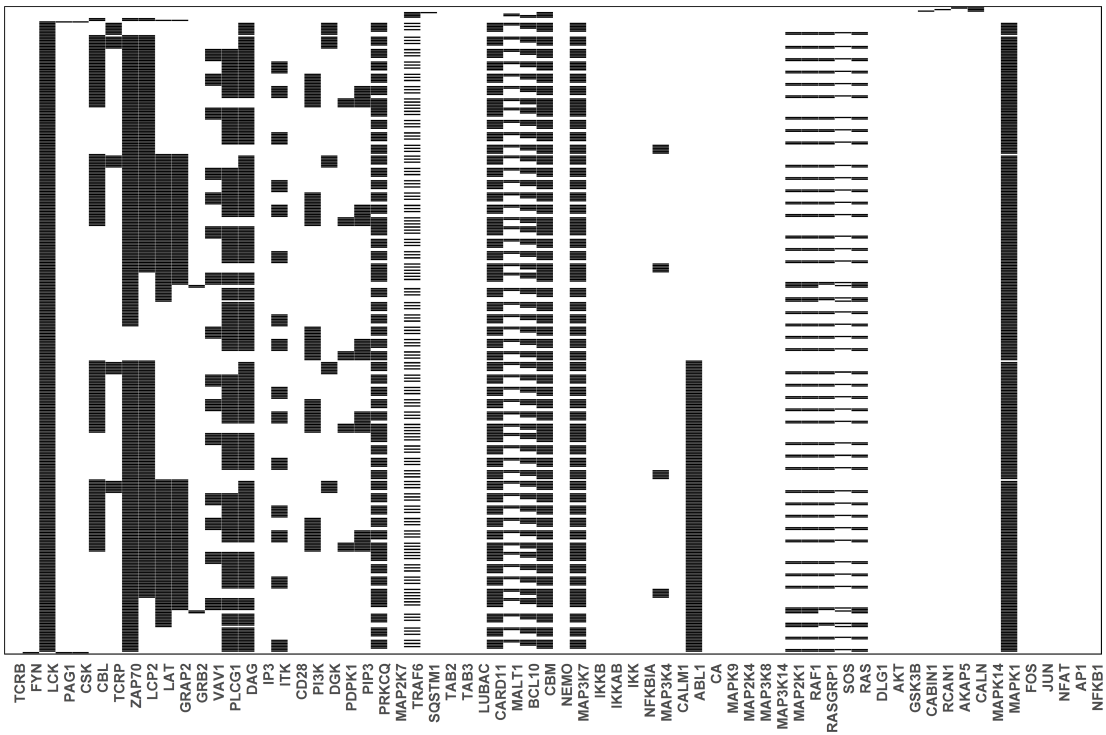


Fig 4. Feedback loops or cycles in the interaction graph. Signaling paths having FBLs from signals 1 and 2 to the major transcription factors identified from the interaction graph. The columns represent the Boolean update equations and are labeled with the updated protein. Each row represents an FBL, and consists of the proteins located along it. On each row, cells with a black background indicate proteins that are along the FBL. There are 419 loops, containing on average 14 proteins.

https://doi.org/10.1371/journal.pone.0176500.g004

PID failure analysis

To study the effects of disease-causing variations on the long-term dynamics of naïve CD4+ T-cells, we perturbed PID proteins in the network model and simulated their dynamics with the normalized HillCube update approach. Twelve PIDs are known to affect the proteins in the network including BCL10, CARD11, IKKB, ITK, LCK, MALT1, MAP3K14, NEMO, NFKBIA, PI3K, PTPRC, TCR (TRAC, a component of the TCR complex) and ZAP70. The proteins were identified from the ImmunoDeficiency Resource [25], IDbases [8], the most recent classification by the IUIS expert committee for PIDs [9] and a recent review [26]. These proteins are expressed at the pre-CD4+ stage during T-cell development and differentiation. The effects of knockouts or overexpression of these proteins to the signaling pathways were investigated by turning them off (on) during simulation. The resulting perturbed attractors were probed for differences compared to the wild type attractor.

The three major TF pathways were dysregulated in the attractors of PIDs involved in the early events of the TCR-dependent T-cell activation including ITK, LCK, PTPRC, TCR and ZAP70 perturbations (Fig 6). AP1 was inactive in the PID attractors for BCL10, CARD11,

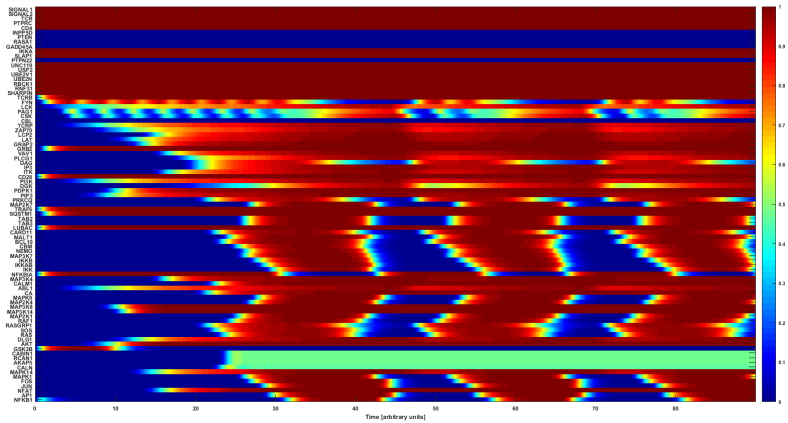


Fig 5. Attractor basin of the CD4+ T-cell network model normalized HillCube simulation. The basin of attractors of the CD4+ T-cell network model simulated using the normalized HillCube algorithm. The horizontal axis denotes time in arbitrary units.

https://doi.org/10.1371/journal.pone.0176500.g005

ITK, LCK, MALT1, MAP3K14 and PTPRC. The NF-κB pathway was dysregulated in all the PID-perturbed attractors, except that for PI3K and NFBIA. The NFKBIA knockout and the PI3K overexpression simulations were identical to the wild type. The perturbations indicate profound effects in the networks for almost all the PIDs. Several novel proteins were found to be affected by the complete and partial knockouts (knockins) of the PID proteins.

Correlation to PID severity

The severity of PIDs varies greatly from very mild to life-threatening conditions. Severe combined immunodeficiency (SCID) is associated with high susceptibility to bacterial, viral and

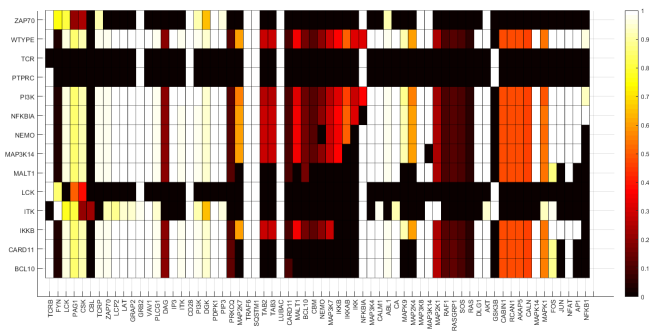


Fig 6. Wild type and PID attractors of the CD4+ T-cell network simulation. The node states for the wild type and the PID-perturbed attractors (knockout perturbation of LCK, ZAP70, ITK, IKKB, NEMO, CARD11, MALT1, BCL10, NFKBIA, PTPRC, MAP3K14 and knockin perturbation of PI3K) attractors. The attractors are represented by the rows while the states of the nodes in the attractors are represented on the columns. The state of a node for an attractor is represented by the color of the cell on the row of the attractor; black means inactive whereas white means activate.

https://doi.org/10.1371/journal.pone.0176500.g006

fungal infections [27]. Persistent infections with respiratory and gastrointestinal viruses and opportunistic pathogens are frequent and often associated with protracted diarrhea and failure to thrive. According to the IUIS classification [9], most of the PIDs in this study are associated with SCIDs with reduced numbers or absent T and B cells. These include BCL10, CARD11, IKKB, ITK, LCK, MALT1, MAP3K14, NEMO, PTPRC, TCR and ZAP70 deficiencies. Interestingly, the attractors for these proteins show severe dysregulation (Fig 6).

Gain-of-function variants in the *PIK3CD* gene, a catalytic subunit of the PI3K heterodimeric complex is associated with a milder PID [28–30]. Additionally, variants in the gene that code for NFKBIA are associated with various forms of ectodermal dysplasia with immunodeficiency (EDA-ID) [31–34]. The attractors for these two proteins are very similar to the wild type form.

Novel PID-associated proteins

The discovery and cataloging of the PIDs is an ongoing effort. With the improvement, development and reduction in the cost of new technologies, more PIDs are identified. Due to the large number, rarity and overlapping symptoms of PIDs, the diagnosis may be late, difficult and costly. Several efforts have been made to ease diagnosis by classifying PIDs [9, 10], predicting and prioritizing candidate genes and proteins [35–38]. The FBLs of our model and the PID-perturbed attractors from simulations provide information about proteins that affect several pathways and could be involved with PIDs. Proteins which are along at least 20 FBLs include the majority of the investigated PIDs and several proteins essential for CD4+ T cell activation and functions. Interestingly, most of these proteins also indicate abrogated signaling in the attractors for most of the PIDs. To evaluate, *in silico*, the effects of perturbing the non-PIDs in Table 1, we performed knockout simulations for each node, except for CBL for which knockin simulation was performed, as CBL is turned off in the wild type attractor. Twenty-one (70%) of the perturbed nodes are impaired in TCR-dependent T cell activation. Further, we investigated the Human Gene Connectome (HGC) (ref) and found that many of the proteins involved in numerous FBLs have significant connections to known PID proteins. Taken together, the genes coding for these proteins are worth considering when prioritizing genes during challenging diagnosis.

Discussion

In this study, we used the normalized HillCube approach to simulate the PID knockout effects in the naïve CD4+ T-cell network dynamics. To achieve this, a network was reconstructed based on evidence from the literature and a previously identified core T-cell network. By using normalized HillCube simulations, we refined and *in silico* validated the reconstructed network. The normalized HillCube perturbation studies qualitatively replicated complete loss-of-function variation effects for several PIDs at CD4+ T-cell developmental stages.

Comparison of the wild type to the PID attractors highlighted significant differences in the signal transduction patterns for ITK, LCK, PTPRC, TCR and ZAP70. The effects of the LCK, PTPRC, TCR and ZAP70 perturbations are severe. Knockout simulations for these proteins qualitatively capture major changes in signaling patterns. The differences between the wild type and MAP3K14, NEMO and IKKB PID simulations were somewhat minor. In the BCL10, MALT1, CARD11, MAP3K14, NEMO and IKKB knockouts, the NF-κB pathway was the most affected. This is because these proteins connect receptor-dependent signals to the distal NF-κB pathway [24]. Knockout of any of these genes may cause the IKK complex, the major NF-κB regulator, to be impaired, leaving NFKBIA bound to NFKB1, and preventing its nuclear transportation and function as a TF [24]. These results show that our approach of simulating effects

Table 1. Number of FBLs along which each protein is in the T cell network model.

Protein	No of FBLs	Effect on NFAT pathway	Effect on NF-κB pathway	Effect on AP1 pathway	BRP	Core proteins[b]
LCK	409	0	0	0		
MAPK1	404	1	1	1	0.00275	NFKBIA (PI3K, PTPRC)
ZAP70	380	0	0	0		
DAG	344	1	0	0		
CBM[a]	316	1	0	0		
PRKCQ	312	1	0	0	0.00024	IKKB (CARD11, LCK, MALT1, NEMO, PTPRC, ITK, PI3K, NFKBIA)
CARD11	312	1	0	0		
MAP3K7	312	1	0	0	0.00281	IKKB (CARD11, MALT1, NEMO, NFKBIA, ZAP70, ITK, PI3K)
LCP2	310	1	0	0	0.00048	ITK (LCK, ZAP70, PTPRC, PI3K, NFKBIA, IKKB)
PLCG1	304	1	0	0	0.00167	ITK (PI3K, LCK, ZAP70)
BCL10	210	1	0	0		
LAT	193	1	0	0	0.00119	ZAP70 (PTPRC, ITK, PI3K, LCK, NFKBIA)
CBL	190	0	0	0	0.00114	ITK (PI3K, ZAP70, LCK, PTPRC)
ABL1	189	0	0	0	0.00633	NFKBIA (LCK, ZAP70, ITK, PI3K)
GRAP2	171	1	0	0	0.00048	ITK (PTPRC, PI3K, NFKBIA, IKKB, MALT1)
TRAF6	160	1	0	0	0.00191	NFKBIA (MALT1, IKKB, CARD11, LCK, NEMO, ZAP70, PI3K)
VAV1	120	1	0	0	0.00036	ITK(PI3K, LCK, ZAP70, NFKBIA)
ITK	120	0	0	0		
PI3K	110	1	1	1		
MALT1	106	1	0	0		
MAP2K1	92	1	1	1	0.00072	PI3K (ITK, NFKBIA)
RAF1	92	1	1	1	0.00329	PI3K (LCK, ITK, NFKBIA)
RAS	92	1	1	1	0.00335	LCK (ZAP70, PI3K, NFKBIA, IKKB)
RASGRP1	86	1	1	1	0.00125	ZAP70 (PI3K, IKKB, MALT1, CARD11, LCK, ITK, PTPRC, NEMO)
PIP3	70	1	0	0		
SOS	47	1	1	1	0.00036	ITK (PI3K, LCK, ZAP70, CARD11)
TCRP	40	1	1	1		
DGK	40	1	1	1	0.00556	NFKBIA (ZAP70, CARD11)
PDPK1	30	1	0	0	0.00102	MALT1 (CARD11, PI3K, IKKB, NFKBIA, LCK, NEMO, ZAP70)
MAP3K4	24	1	0	0	0.01673	PI3K (LCK, IKKB, NEMO)

Rows with tan background are for PIDs.

[a]CBM deficiency is considered as a PID because it is a complex, all of whose components are related to PIDs.

[b]The first core protein is the most significant to the target and those in parenthesis are other significant ones for the target (BRP < 0.05).

https://doi.org/10.1371/journal.pone.0176500.t001

of protein variations in networks is effective when the affected proteins are in the core of the interconnected network or along non-redundant paths belonging to crucial pathways. No major changes were revealed in overexpression perturbation, as in PI3K, or redundant signaling path between the receptor to the TF. In the NFKBIA PID heterozygous variants that either lack the phosphorylation sites [31, 32] or truncate the protein [34] protect it from phosphorylation-induced proteosomal degradation. The inactivated NFKBIA sequesters NFKB1 in the cytosol [32]. Thus, the deficient NFKBIA acts as a dominant negative form for NFKB1, reducing NFKB1's activity and causing the reduction of TCR activation-dependent cytokine

response [32]. Indeed, there were no observed differences between the PI3K and NFKBIA PID simulations compared to the wild type. LCK and ZAP70 perturbations that cause major effects are present in over 90% of the FBLs in the network and CARD11 and the CBM complex in 75% of the FBLs. Seven of the 12 PID proteins emerge in the FBLs, most of which are proximal TCR activation events, highlighting the fact that the simulation studies are effective for detecting effects of centrally located proteins.

Antigen-TCR complex ligation causes conformational alterations of CD3 chains, which contain immunoreceptor tyrosine-based activation motifs (ITAMs) on which they are phosphorylated by LCK [3]. This is an essential step in early TCR activation. The LCK kinase activity is regulated by the antagonistic actions of the membrane protein tyrosine phosphatase PTPRC and the carboxy-terminal Src kinase (CSK) (30). The phosphorylation of Tyr505 in LCK by CSK inhibits LCK activity via auto-phosphorylation of Tyr394 in the catalytic domain. The dephosphorylation of the Tyr505 by PTPRC relieves this inhibition [39]. TCR is crucial for T cell activation and cytokine response, and simulation of TCR deficiency shows profound impairment of all TF pathways. A homozygous variant of TRAC, a crucial component of the TCR complex, causes this deficiency [40]. The deficiency is associated with lymphadenopathy, recurrent infections and hepatosplenomegaly. Because the increased activity of LCK is crucial for the T cell response after antigen stimulation, the PTPRC knockout causes a severe perturbation. This is confirmed by disease-causing variations in the gene [41–45]. The known variants include large deletions [44] and amino acid substitutions [45]. Immunodeficiencies caused by the lack of LCK activity lead to T cells that are low in number and non-responsive, which in turn causes susceptibility to infections. Our PTPRC-perturbed simulations indicate that all the signaling paths of NFAT, NFKB1 and AP1 TFs, crucial for TCR-dependent response, are disrupted.

The activation of LCK is a crucial early step for T-cell activation and response. The phosphorylation of the CD3 ITAMs leads to the recruitment of ZAP70 and its activation by LCK. ZAP70 subsequently phosphorylates LAT, leading to the formation of the LAT signalosome (the proximal signaling complex) [46]. LAT signalosome transduces signals to pathways that are indispensable for the three major TFs necessary for T-cell activation and response. Thus, the improper constitution of this signaling site affects multiple pathways and disrupts the transduction of TCR activation signals, as verified by our simulations.

The absence of LCK signaling disrupts the NFAT pathways and abrogates the T-cell response. The LCK deficiency is associated to naive CD4+ T-cell lymphopenia, respiratory tract infections, and early-onset autoimmune inflammation [47–49]. The major effects of this PID on naïve CD4+ T-cells are a profoundly defective TCR signaling, lack of calcium/magnesium signaling and defective NF-κB response. Our simulation of the knockout perturbation confirms the dysregulation of most signaling events associated with the calcium signaling, thereby affecting the AP1, NFKB1 and NFAT signaling pathways. LCK and ZAP70, the two vital components necessary for the formation of the LAT signalosome, are turned off in the LCK-perturbed attractor. This suggests that the LAT signalosome is disrupted and thus, downstream signaling is impaired. As shown in Fig 6, the signaling components required for the AP1, NF-κB, NFAT family of proteins, including the calcium-dependent signaling, are turned off in the LCK knockout attractor. The affected signaling components include PLCG2, PIP2, IP3, DAG and CALN.

Because of the proximity of ZAP70 to LCK in the early activated TCR signaling events, the effects of ZAP70 are expected to be similar. This is indeed the case. Partially affected signaling occurs in ZAP70 deficiency, but downstream responses, like proliferation, are abrogated because of the TCR signaling defect. Severe conditions caused by the ZAP70 deficiency have been diagnosed in several patients [50–55]. Like the PTPRC and LCK knockout simulations,

the major effectors associated to the calcium signaling are turned off in the ZAP70 perturbed attractor. Based on these results, the activated T-cells would become anergic and/or undergo apoptosis. SYK, the ZAP70 homolog in non-T-cells, is expressed at high levels in the CD4+ T-cells of ZAP70-deficient patients [50, 53]. The SYK expression might compensate for the lack of ZAP70, and has been used to explain the less severe phenotype of the ZAP70 deficiency [50].

During the constitution of the LAT signalosome, LCP2 and PLCG1 bind to LAT and are phosphorylated by ZAP70 [46]. The phosphorylated LCP2 then recruits ITK, which leads to the activation of PLCG1. PLCG1 hydrolyzes its substrate PIP2 to generate second messengers, IP3 and DAG. ITK is a non-receptor tyrosine kinase expressed in T-cells and has been described as an important component of proximal TCR signaling [56].

Several homozygous ITK variants cause PID [57–60]. The ITK deficiency is associated with naive CD4+ T-cell lymphopenia, modest change in the number of CD4+ T-cells, impaired positive and negative selection of thymocytes due to reduced TCR signal levels, recurrent infections (for example, herpes virus infections), autoimmune cytopenias, lymphoproliferation, lymphadenopathy and hepatosplenomegaly. Genotype studies point to a twofold increase in activated CD4+ T-cells, impaired activation-induced cell death and decreased levels of TCR signaling. Additionally, there is evidence that TXK could substitute for ITK [61]. The lack of ITK in mice is mitigated by the ability of TXK to activate PLCG1 [62]. ITK is present in both the strongly connected component and several (29%) of the FBLs. These findings indicate that in the absence of ITK, T-cells are activated, but signaling resulting from TCR stimulation leads to impaired response. Indeed, the attractors from our perturbed simulations showed abrogation of the NFAT, AP1 and NF-κB pathways. This agrees with normal, but progressive decrease in T cell numbers that may be caused by defective response in the TCR-dependent response pathways, which are indispensable for IL-2 transactivation and T cell response [63, 64].

The constitution of the CBM complex is an essential event in the regulation of NF-κB pathway. After the TCR/CD28 activation, PRKCQ is activated and recruited to the proximal signalosome. Here, PRKCQ activates CARD11 [65], which leads to its association with BCL10. Because BCL10 is constitutively bound to MALT1, the association of CARD11 to BCL10 leads to the formation of the CBM complex. Several PIDs have been connected to variations that occur on the genes that code for CARD11 [66], MALT1 [67] and BCL10 [68]. The CARD11 PID case is caused by a homozygous premature stop codon on the gene that codes for CARD11, and truncates its kinase-like domain. A homozygous variant in the CARD domain of *MALT1* causes MALT1 PID. The known BCL10 PID case is due to a homozygous splice-site variation at intron 1 of the gene encoding BCL10. The CARD11 PID is associated with hypogammaglobulinemia, severe interstitial pneumonia, dyspnea and respiratory tract infections [66]. The MALT1 deficiency is associated with bronchiectasis, mastoiditis, chronic aphthous ulcers, gastritis, gingivitis, duodenitis and meningitis while the BCL10 PID is associated with hypogammaglobulinemia, gastroenteritis, otitis, respiratory tract infection and several viral infections [68]. Although the CBM PIDs show normal T cell counts, the BCL10 and MALT1 deficiencies show predominantly naïve CD4+ T cells, including severely abrogated TCR-dependent NF-kB signaling and cytokine response [69]. As expected, the pathways for NF-κB and AP1 are severely disrupted in the attractors of the CARD11, MALT1 and BCL10 PIDs.

The major regulator of NF-κB is the IKK complex [24]. It consists of two protein kinases, IKKA and IKKB and a regulatory protein, NEMO [70]. The activation of the IKK complex is NEMO-dependent. After the TCR/CD28 activation PRKCQ is activated and recruited to the proximal signalosome, where it activates CARD11 [65], which leads to the formation of the CBM complex. The TRAF6 oligomerizes with the CBM complex through the association with

MALT1 and BCL10 [71]. This oligomerization recruits UBE2V1 which polyubiquitinates and, thus, activates TRAF6 [72]. The activated TRAF6 in turn activates MAP3K7, which subsequently coordinates the assembly of the IKK complex [71, 73].

Some PIDs have been linked to both IKKB and NEMO [74–77]. A complete loss of function homozygous truncating variant, a duplicating variant, and a nonsynonymous nucleotide substitution on the gene that codes for IKKB have been reported to cause the disease [77–79]. IKKB deficiency is associated with life-threatening bacterial, fungal, and viral infections, defective immunoglobulin production and hypo- or agammaglobulinemia. Although T cell numbers are normal, T cell subsets are lower, and peripheral T cells fail to respond to stimulation. *IKBKB* loss of function variants abrogates signaling and response via the NF-κB pathway in these patients [24]. Genetic studies have revealed several PID cases linked to *IKBKG*, the gene that codes for NEMO [74, 75, 80–82]. The disease results from amino acid substitution and exon skipping variations. The NEMO deficiency is associated with anhidrotic ectodermal dysplasia, polysaccharide non-response, various infectious diseases, colitis, ectodermal dysplasia, conical teeth, variable defects of skin pigmentation and monocyte dysfunction [74, 75]. The T cell counts are normal but TCR activation is impaired, especially NF-κB activation. In accordance with these studies, our simulation indicates that the NEMO and IKKB perturbations lead to inactivation of NF-κB, despite normal activation of AP1 and NFAT [24, 83, 84].

MAP3K14 is a member of the family of mitogen-activated protein kinases that is involved in both the canonical [24] and non-canonical [85] NF-κB pathways. In the canonical NF-κB pathway, the CD28 co-stimulatory signal is required for the MAP3K14 activation through MAP3K8 (COT). After activation by AKT1, MAP3K8 activates MAP3K14, which in turn contributes in the activation and subsequent ubiquitination of NFKBIA [71, 73, 86]. The ubiquitination of NFKBIA releases NFKB1 which is translocated into the nucleus and results in T-cell response. In the non-canonical NF-κB pathway, MAP3K14 associates with IKKA to induce the phosphorylation and subsequent ubiquitination of the p100 subunit [85, 87]. This leads to the proteolysis of NFKB2/p100 to NFKB2/p52-RELB dimer, which is translocated to the nucleus and transactivates κB-containing genes for response [85].

A PID caused by a biallelic variation in the gene coding for MAP3K14 protein leads to loss of its kinase activity [88]. This variant disrupts both the canonical and non-canonical NF-κB pathways in immune response cell-types [88]. Despite the normal overall T cell numbers, several T cell subsets show defective response and perturbation. The MAP3K14 PID is associated with several microbial infections, including bacterial and viral infections [88]. The MAP3K14 PID-perturbed simulations are in accordance with its crucial and non-redundant role in T cells as seen in the defective activation of NFKB1, albeit normal activation of AP1, NFAT and MAPK14 [89].

The results for simulations of NFKBIA and PI3K did not differ from wild type. To investigate the effects of variants and knockouts in these proteins, dedicated networks would be needed with more information about downstream factors.

Our results show PID-caused trends in the cellular dynamics of the CD4+ T-cells when the affected proteins are involved in non-redundant paths along major TF signaling pathways. The downstream signaling events show minor effect on the network dynamics than the early events. This paper is the first attempt, as far as we are aware, to investigate, with systems biological simulations, the effects of variations in immune response proteins in PIDs. We found profound effects in the ITK, LCK, PTPRC, TCR and ZAP70 perturbed simulations, and less profound but noticeable effects in the BCL10, CARD11, IKKB, MALT1, MAP3K14 and NEMO perturbed simulations.

The non-PID proteins in Table 2 are indispensable for T cell activation and response, are affected in several of the simulated PID attractors and have also been associated with other

Table 2. Tuned parameters of nodes in the Odefy-simulated T cell network model.

Influenced node	Influencing node(s)	τ	n	k
PAG1	[]a	1	20	0.9
PAG1	[]	1	20	0.9
DAG	DGK	1	20	0.9
DGK	[]	1	20	0.9
DGK	[]	1	3	0.9
DGK	[]	1	3	0.9
LCK	MAPK1	10	20	0.1
CBL	[]	3	20	0.9
CALN	CABIN1	1	3	0.9
CALN	RCAN1	1	3	0.9
CALN	AKAP5	1	3	0.9

[a]All influencing nodes.

PAG1, phosphoprotein membrane anchor with glycosphingolipid microdomains 1; DAG, second messenger, diacylglycerol; DGK, diacylglycerol kinases; LCK, LCK proto-oncogene, Src family tyrosine kinase; MAPK1, mitogen-activated protein kinase 1 (ERK); CBL, Cbl proto-oncogene; CALN, calcineurin complex; CABIN1, calcineurin Binding Protein 1, RCAN1, regulator of calcineurin 1, AKAP5, A-kinase anchoring protein 5.

https://doi.org/10.1371/journal.pone.0176500.t002

diseases. Several of them have been identified as candidate PIDs. *VAV1*, *RAF1*, *LAT*, *LCP2* and *MAPK1* were identified as candidate PID genes with high confidence by Keerthikumar et al. [37]. Moreover, 15 out of 22 of the proteins are predicted to be candidates by another recent study [38]. These include LCP2, CBL, TRAF6, MAP3K7, VAV1, PLCG1, PRKCQ, RAF1, ABL1, PDPK1, GRAP2, LAT, MAPK1, MAP3K4 and MAP2K1. Several of the candidate genes are central in the Human Gene Connectome (Table 2) providing independent proof for their significance. As the connectome is not complete, the fact that there is no support from this method does not mean that our findings were not significant even from this point of view.

Nine of the proteins in Table 2 are protein kinases (MAPK1, PRKCQ, MAP3K7, PLCG1, LAT, MAP2K1, RAF1, PDPK1, MAP3K4), 4 are mitogen-activated protein kinases (MAPK1, MAP2K1, RAF1 and MAP3K4), 3 are serine-threonine kinases (PRKCQ, MAP3K7, and PDPK1), and 3 have guanyl-nucleotide exchange factor activity (VAV1, RASGRP1 and SOS). Four of the proteins are linked to various forms of the Noonan syndrome (CBL, MAP2K1, RAF1 and SOS), 5 to various types of tumors (MAPK1, PRKCQ, ABL1, GRAP2 and RAS) and one to an autoimmune disorder (RASGRP1). Seven of the genes are not linked to any disease (MAP3K7, LCP2, LAT, VAV1, DGK, PDPK1 and MAP3K4). The listed proteins are strong PID candidates; however, their involvement in PIDs needs to be experimentally verified. In the case of the NFKBIA perturbed simulations we observed local effect on NF-κB and for PI3K, no effects. Further simulation studies of these PIDs will require more specific networks, if applicable.

Several studies suggest candidate PID genes [35, 37, 38]. Ortutay and Vihinen constructed a PPI network of immune system-specific proteins, proteins with high network statistics and PID-related Gene Ontology term enrichment scores [35]. Itan and Casanova identified the top 1% of genes that were biologically close to known PIDs and, and from these selected the ones with similar Gene Ontology terms as the known PIDs [38]. A machine learning technique, support vector machine, was applied by Keerthikumar and colleagues to identify candidate

PIDs by utilizing binary features from PIDs and non-PIDs [37]. The above approaches were successful in identifying several candidate genes that were subsequently verified to be PID related. Our approach focuses on T-cell-specific PIDs and how they affect other components of the cellular signaling dynamics. This, as well as other evidence presented above, allowed us to identify the candidate PIDs.

Diagnosis and prognosis of PIDs is still often problematic. Our approach provides novel insights into the mechanisms of PID effects on signaling cascades and may highlight novel targets for therapy downstream of the defective proteins. The presented approach can be used to study PIDs of any cellular system and even diseases outside the immune system.

Methods

Network reconstruction and analysis

The T-cell PPI network (TPPIN), a core network of PPIs specific to T-cells [17], was used as the basis for extensive literature survey and the reconstruction of the Boolean equations for the T cell model. Only those nodes that have been demonstrated to play a crucial role in the TCR/CD28-dependent activation of CD4+ T cells were retained.

The CellNetAnalyzer version 2016.1 [18] was used for identifying feedback loops in the underlying interaction graph of the model. The base R software version 3.2.3 [90] and Cytoscape version 3.3.0 [91] were used for data analysis and network visualization, respectively. The strongly connected components were calculated using *igraph*, a library for network and graph analyses in R [92].

A Boolean model consists of N nodes/proteins $X_1, X_1, \ldots, X_N$. The proteins are represented by variables x_i that take values {0, 1} [93]. Each protein, x_i is influenced by a set of proteins $R_i = \{X_1, X_1, \ldots, X_N\}$ connected to it. Based on the values of their influencing proteins R_i, for each time step, the value of each protein x_i, is calculated from the update function $B:\{0, 1^N\}$. Because the time is discretized in Boolean simulations, at time point $t + 1$, updates are done synchronously as follows [93, 94],

$$x_i(t+1) = B_i(x_{i1}(t),\ x_{i2}(t), \ldots,\ x_{iN_i}(t)) \in \{0,\ 1\},\ i = 1,\ 2,\ \ldots,\ N.$$

The Boolean update functions, B_i, are converted into a system of continuous ordinary differential equation (ODE) model where x_i takes values [0, 1] using the following ODE equation

$$\dot{x}_i = \frac{1}{\tau_i}(\bar{B}_i\ (\bar{x}_{i1},\ \bar{x}_{i2}, \ldots,\ \bar{x}_{iN_i}) - \ \bar{x}_i),$$

where, $\bar{B}_i$ is a continuous homologue of the discrete function B_i, parameter τ_i represents the life-time of the protein, and $\bar{x}_i$ describes its decay.

Odefy [16], a toolbox compatible with MATLAB, transforms B_i to the ODE system and computes the solution of the system using the BooleCubes [95] as follows,

$$\bar{B}^I(\bar{x}_1,\ \bar{x}_2, \ldots,\ \bar{x}_N) = \sum_{x_1=0}^{1}\sum_{x_1=0}^{1}\cdots\sum_{x_N=0}^{1}\left[B(x_1,\ x_2,\ \ldots,\ x_N).\prod_{i=1}^{N}(x_i\bar{x}_i\ + (1-x_i)(1-\bar{x}_i))\right].$$

$\bar{B}^I$, the BooleCube, is obtained from the multilinear interpolation of the Boolean update function B_i. Biomolecular interactions show switch-like behavior and are modeled using sigmoidal functions. Thus, the Hill function, $f(\bar{x}) = \bar{x}^n/(\bar{x}^n + k^n)$, was used to smoothen the affine multilinear BooleCube, to obtain the sinusoidal HillCube [95]. Hence, the parameter n was introduced (the Hill coefficient or slope of the Hill function), to represent the cooperativity between the protein interactions and parameter k represent the value at which the

activation is half-maximal. The HillCube is obtained from the BooleCube as follows,

$$\bar{B}^H(\bar{x}_1, \ldots, \bar{x}_N) = \bar{B}^I(f_1(\bar{x}_1), \ldots, f_N(\bar{x}_N)).$$

To obtain perfect homologues of the Boolean update functions B_i, the HillCube functions are normalized to the unit interval to give the normalized HillCube [95] as follows,

$$\bar{B}^{Hn}(\bar{x}_1, \ldots, \bar{x}_N) = \bar{B}^I\left(\frac{f_1(\bar{x}_1)}{f_1(1)}, \ldots, \frac{f_N(\bar{x}_N)}{f_N(1)}\right).$$

The network model used in this study is available in SBML qual format (S1 Text) on the website http://structure.bmc.lu.se/tcell_net/web_session/#/.

Basin of attraction and attractor identification

The Odefy was used to simulate the qualitative dynamics of the network model. It provides simulation algorithms for both synchronous and asynchronous updates and allows simulations based on the BooleCubes [16]. We used the normalized HillCube functions, which represent the normalized BooleCubes in the range [0–1]. Boolean dynamic simulations were performed using normalized HillCube simulations [95]. Except for nodes involved in some negative feedback loops, the default parameter values were used. The default parameters for the normalized HillCube were $n = 3$, $k = 0.5$ and $\tau = 1$. Table 2 lists non-default parameters for nodes on some feedback loops. The variable n represents the Hill exponent of the Hill function and is used for converting the discrete Boolean update functions that take value {0, 1} into their continuous BooleCube equivalents that have values [0, 1]. It captures the influence that nodes of the same Boolean equation have on each other. k is a variable to control the continuous relaxation of the Boolean step function. It represents the value at half-maximal activation of a protein. τ is a decay parameter; for each protein, the higher its value the slower the decay of the protein. The simulations were run until the network dynamics settled in an attractor.

Perturbation

The Analysis of PID effects was performed for each protein encoded by a PID gene using the normalized HillCube simulations. For each perturbation, the node was converted to an input before assigning a state, either off or on, depending on the PID. For example, if the PID occurs as a result of over-activity of the protein, then the perturbed state is ON. This state was maintained until the simulation transitioned into the attractor. The parameter values used in the wild type simulations were maintained for all the PID perturbed simulations. The end result of the simulation represents the perturbed PID attractor.

Primary immunodeficiency data

PID proteins expressed after the pre-CD4+ T-cell development stage were retrieved from the IDbases [8], the most recent updated IUIS expert committee classification of PID data [9], and a recent survey [26], and used for the PID failure mode simulations. The PIDs included LCK, ZAP70, ITK, IKKB, NEMO, CARD11, MALT1, BCL10, NFKBIA, PTPRC, MAP3K14 and PI3K deficiencies.

Supporting information

S1 Table. CD4+ T-cell activation Boolean network model update equations. The table lists Boolean equations of protein activation used in the network model and simulations.
(DOCX)

S1 Text. SBML qual. The CD4+ T-cell network model in SBML qual format. The contains the CD4+ T-cell network qualitative model in the SBML qual format.
(SBML)

Author Contributions

Conceptualization: VM GNT.

Data curation: GNT.

Formal analysis: GNT.

Funding acquisition: VM.

Investigation: GNT VM.

Methodology: GNT VM.

Project administration: GNT VM.

Resources: VM.

Software: GNT.

Supervision: VM.

Validation: GNT.

Visualization: GNT VM.

Writing – original draft: GNT.

Writing – review & editing: GNT VM.

References

1. Ortutay C, Vihinen M. Immunome: A reference set of genes and proteins for systems biology of the human immune system. Cell Immunol. 2006; 244(2):87–9. https://doi.org/10.1016/j.cellimm.2007.01.012 PMID: 17434156
2. Samarghitean C, Ortutay C, Vihinen M. Systematic classification of primary immunodeficiencies based on clinical, pathological, and laboratory parameters. Journal of immunology (Baltimore, Md: 1950). 2009; 183(11):7569–75.
3. Smith-Garvin JE, Koretzky GA, Jordan MS. T Cell Activation. Annual Review of Immunology. 2009; 27:591–619. https://doi.org/10.1146/annurev.immunol.021908.132706 PMID: 19132916
4. Abbas AK, Murphy KM, Sher A. Functional diversity of helper T lymphocytes. Nature. 1996; 383(6603):787–93. https://doi.org/10.1038/383787a0 PMID: 8893001
5. Nakayamada S, Takahashi H, Kanno Y, O'Shea JJ. Helper T cell diversity and plasticity. Curr Opin Immunol. 2012; 24(3):297–302. https://doi.org/10.1016/j.coi.2012.01.014 PMID: 22341735
6. Saez-Rodriguez J, Simeoni L, Lindquist JA, Hemenway R, Bommhardt U, Arndt B, et al. A logical model provides insights into T cell receptor signaling. PLoS Comput Biol. 2007; 3(8):e163. https://doi.org/10.1371/journal.pcbi.0030163 PMID: 17722974
7. Saadatpour A, Wang RS, Liao A, Liu X, Loughran TP, Albert I, et al. Dynamical and structural analysis of a T cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia. PLoS Comput Biol. 2011; 7(11):e1002267. https://doi.org/10.1371/journal.pcbi.1002267 PMID: 22102804
8. Piirila H, Valiaho J, Vihinen M. Immunodeficiency mutation databases (IDbases). Hum Mutat. 2006; 27(12):1200–8. https://doi.org/10.1002/humu.20405 PMID: 17004234
9. Picard C, Al-Herz W, Bousfiha A, Casanova JL, Chatila T, Conley ME, et al. Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. J Clin Immunol. 2015.

10. Samarghitean C, Ortutay C, Vihinen M. Systematic classification of primary immunodeficiencies based on clinical, pathological, and laboratory parameters. J Immunol. 2009; 183(11):7569–75. https://doi.org/10.4049/jimmunol.0901837 PMID: 19917694

11. del Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. Curr Opin Biotechnol. 2010; 21(4):566–71. https://doi.org/10.1016/j.copbio.2010.07.010 PMID: 20709523

12. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(21):8685–90. https://doi.org/10.1073/pnas.0701361104 PMID: 17502601

13. Li FT, Long T, Lu Y, Ouyang Q, Tang C. The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci U S A. 2004; 101(14):4781–6. https://doi.org/10.1073/pnas.0305937101 PMID: 15037758

14. Thakar J, Saadatpour-Moghaddam A, Harvill ET, Albert R. Constraint-based network model of pathogen-immune system interactions. J R Soc Interface. 2009; 6(36):599–612. https://doi.org/10.1098/rsif.2008.0363 PMID: 18952547

15. von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust development module. Nature. 2000; 406(6792):188–92. https://doi.org/10.1038/35018085 PMID: 10910359

16. Krumsiek J, Poelsterl S, Wittmann DM, Theis FJ. Odefy—From discrete to continuous models. BMC Bioinformatics. 2010; 11:233-. https://doi.org/10.1186/1471-2105-11-233 PMID: 20459647

17. Teku GN, Ortutay C, Vihinen M. Identification of core T cell network based on immunome interactome. BMC Syst Biol. 2014; 8:17-. https://doi.org/10.1186/1752-0509-8-17 PMID: 24528953

18. Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED. A methodology for the structural and functional analysis of signaling and regulatory networks. BMC Bioinformatics. 2006; 7:56. https://doi.org/10.1186/1471-2105-7-56 PMID: 16464248

19. Linsley PS, Bradshaw J, Urnes M, Grosmaire L, Ledbetter JA. CD28 engagement by B7/BB-1 induces transient down-regulation of CD28 synthesis and prolonged unresponsiveness to CD28 signaling. J Immunol. 1993; 150(8 Pt 1):3161–9. PMID: 7682233

20. Sugiyama Y, Kakoi K, Kimura A, Takada I, Kashiwagi I, Wakabayashi Y, et al. Smad2 and Smad3 are redundantly essential for the suppression of iNOS synthesis in macrophages by regulating IRF3 and STAT1 pathways. Int Immunol. 2012; 24(4):253–65. https://doi.org/10.1093/intimm/dxr126 PMID: 22331441

21. Baltanas FC, Perez-Andres M, Ginel-Picardo A, Diaz D, Jimeno D, Liceras-Boillos P, et al. Functional Redundancy of Sos1 and Sos2 for Lymphopoiesis and Organismal Homeostasis and Survival. Mol Cell Biol. 2013; 33(22):4562–78. https://doi.org/10.1128/MCB.01026-13 PMID: 24043312

22. Guo X, Wang XF. Signaling cross-talk between TGF-beta/BMP and other pathways. Cell Res. 2009; 19(1):71–88. https://doi.org/10.1038/cr.2008.302 PMID: 19002158

23. Kannan A, Huang W, Huang F, August A. Signal transduction via the T cell antigen receptor in naive and effector/memory T cells. Int J Biochem Cell Biol. 2012; 44(12):2129–34. https://doi.org/10.1016/j.biocel.2012.08.023 PMID: 22981631

24. Mitchell S, Vargas J, Hoffmann A. Signaling via the NFkappaB system. Wiley Interdiscip Rev Syst Biol Med. 2016; 8(3):227–41. https://doi.org/10.1002/wsbm.1331 PMID: 26990581

25. Samarghitean C, Väliaho J, Vihinen M. IDR knowledge base for primary immunodeficiencies. Immunome Res. 2007; 3:6. https://doi.org/10.1186/1745-7580-3-6 PMID: 17394641

26. Vihinen M. Immunodeficiency, Primary: Affecting the Adaptive Immune System. eLS. Chichester: John Wiley & Sons Ltd, Chichester; 2015.

27. van der Burg M, Gennery AR. The expanding clinical and immunological spectrum of severe combined immunodeficiency. Eur J Pediatr. 2011; 170(5):561–71. https://doi.org/10.1007/s00431-011-1452-3 PMID: 21479529

28. Angulo I, Vadas O, Garcon F, Banham-Hall E, Plagnol V, Leahy TR, et al. Phosphoinositide 3-kinase delta gene mutation predisposes to respiratory infection and airway damage. Science. 2013; 342(6160):866–71. https://doi.org/10.1126/science.1243292 PMID: 24136356

29. Lucas CL, Kuehn HS, Zhao F, Niemela JE, Deenick EK, Palendira U, et al. Dominant-activating germline mutations in the gene encoding the PI(3)K catalytic subunit p110delta result in T cell senescence and human immunodeficiency. Nat Immunol. 2014; 15(1):88–97. https://doi.org/10.1038/ni.2771 PMID: 24165795

30. Crank MC, Grossman JK, Moir S, Pittaluga S, Buckner CM, Kardava L, et al. Mutations in PIK3CD can cause hyper IgM syndrome (HIGM) associated with increased cancer susceptibility. J Clin Immunol. 2014; 34(3):272–6. https://doi.org/10.1007/s10875-014-0012-9 PMID: 24610295

31. McDonald DR, Mooster JL, Reddy M, Bawle E, Secord E, Geha RS. Heterozygous N-terminal deletion of IkappaBalpha results in functional nuclear factor kappaB haploinsufficiency, ectodermal dysplasia,

and immune deficiency. J Allergy Clin Immunol. 2007; 120(4):900–7. https://doi.org/10.1016/j.jaci.2007.08.035 PMID: 17931563

32. Courtois G, Smahi A, Reichenbach J, Doffinger R, Cancrini C, Bonnet M, et al. A hypermorphic Ikappa-Balpha mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. J Clin Invest. 2003; 112(7):1108–15. https://doi.org/10.1172/JCI18714 PMID: 14523047

33. Janssen R, van Wengen A, Hoeve MA, ten Dam M, van der Burg M, van Dongen J, et al. The same IkappaBalpha mutation in two related individuals leads to completely different clinical syndromes. J Exp Med. 2004; 200(5):559–68. https://doi.org/10.1084/jem.20040773 PMID: 15337789

34. Lopez-Granados E, Keenan JE, Kinney MC, Leo H, Jain N, Ma CA, et al. A novel mutation in NFKBIA/IKBA results in a degradation-resistant N-truncated protein and is associated with ectodermal dysplasia with immunodeficiency. Hum Mutat. 2008; 29(6):861–8. https://doi.org/10.1002/humu.20740 PMID: 18412279

35. Ortutay C, Vihinen M. Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res. 2009; 37(2):622–8. https://doi.org/10.1093/nar/gkn982 PMID: 19073697

36. Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 2012; 279(5):678–96. https://doi.org/10.1111/j.1742-4658.2012.08471.x PMID: 22221742

37. Keerthikumar S, Bhadra S, Kandasamy K, Raju R, Ramachandra YL, Bhattacharyya C, et al. Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. DNA Res. 2009; 16(6):345–51. https://doi.org/10.1093/dnares/dsp019 PMID: 19801557

38. Itan Y, Casanova JL. Novel primary immunodeficiency candidate genes predicted by the human gene connectome. Front Immunol. 2015; 6:142. https://doi.org/10.3389/fimmu.2015.00142 PMID: 25883595

39. Thomas ML, Brown EJ. Positive and negative regulation of Src-family membrane kinases by CD45. Immunol Today. 1999; 20(9):406–11. PMID: 10462740

40. Morgan NV, Goddard S, Cardno TS, McDonald D, Rahman F, Barge D, et al. Mutation in the TCRalpha subunit constant gene (TRAC) leads to a human immunodeficiency disorder characterized by a lack of TCRalphabeta+ T cells. J Clin Invest. 2011; 121(2):695–702. https://doi.org/10.1172/JCI41931 PMID: 21206088

41. Cale CM, Klein NJ, Novelli V, Veys P, Jones AM, Morgan G. Severe combined immunodeficiency with abnormalities in expression of the common leucocyte antigen, CD45. Arch Dis Child. 1997; 76(2):163–4. PMID: 9068311

42. Hermiston ML, Xu Z, Weiss A. CD45: A critical regulator of signaling thresholds in immune cells. Annu Rev Immunol. 2003; Volume 21:107–37.

43. Kung C, Pingel JT, Heikinheimo M, Klemola T, Varkila K, Yoo LI, et al. Mutations in the tyrosine phosphatase CD45 gene in a child with severe combined immunodeficiency disease. Nat Med. 2000; 6(3):343–5. https://doi.org/10.1038/73208 PMID: 10700239

44. Roberts JL, Buckley RH, Luo B, Pei J, Lapidus A, Peri S, et al. CD45-deficient severe combined immunodeficiency caused by uniparental disomy. Proc Natl Acad Sci U S A. 2012; 109(26):10456–61. https://doi.org/10.1073/pnas.1202249109 PMID: 22689986

45. Tchilian EZ, Wallace DL, Wells RS, Flower DR, Morgan G, Beverley PC. A deletion in the gene encoding the CD45 antigen in a patient with SCID. J Immunol. 2001; 166(2):1308–13. PMID: 11145714

46. Werlen G, Palmer E. The TCR signalosome: a dynamic structure with expanding complexity. Curr Opin Immunol. 2002; 14(3):299–305. PMID: 11973126

47. Gibson S, Truitt K, Lu Y, Lapushin R, Khan H, Imboden JB, et al. Efficient CD28 signalling leads to increases in the kinase activities of the TEC family tyrosine kinase EMT/ITK/TSK and the SRC family tyrosine kinase LCK. Biochem J. 1998; 330 (Pt 3)(Pt 3):1123–8.

48. Hauck F, Randriamampita C, Martin E, Gerart S, Lambert N, Lim A, et al. Primary T-cell immunodeficiency with immunodysregulation caused by autosomal recessive LCK deficiency. J Allergy Clin Immunol. 2012; 130(5):1144–52 e11. https://doi.org/10.1016/j.jaci.2012.07.029 PMID: 22985903

49. Sawabe T, Horiuchi T, Nakamura M, Tsukamoto H, Nakahara K, Harashima SI, et al. Defect of lck in a patient with common variable immunodeficiency. Int J Mol Med. 2001; 7(6):609–14. PMID: 11351273

50. Toyabe S, Watanabe A, Harada W, Karasawa T, Uchiyama M. Specific immunoglobulin E responses in ZAP-70-deficient patients are mediated by Syk-dependent T-cell receptor signalling. Immunology. 2001; 103(2):164–71. https://doi.org/10.1046/j.1365-2567.2001.01246.x PMID: 11412303

51. Turul T, Tezcan I, Artac H, de Bruin-Versteeg S, Barendregt BH, Reisli I, et al. Clinical heterogeneity can hamper the diagnosis of patients with ZAP70 deficiency. Eur J Pediatr. 2009; 168(1):87–93. https://doi.org/10.1007/s00431-008-0718-x PMID: 18509675

52. Karaca E, Karakoc-Aydiner E, Bayrak OF, Keles S, Sevli S, Barlan IB, et al. Identification of a novel mutation in ZAP70 and prenatal diagnosis in a Turkish family with severe combined immunodeficiency disorder. Gene. 2013; 512(2):189–93. https://doi.org/10.1016/j.gene.2012.10.062 PMID: 23124046

53. Hauck F, Blumenthal B, Fuchs S, Lenoir C, Martin E, Speckmann C, et al. SYK expression endows human ZAP70-deficient CD8 T cells with residual TCR signaling. Clin Immunol. 2015; 161(2):103–9. https://doi.org/10.1016/j.clim.2015.07.002 PMID: 26187144

54. Picard C, Dogniaux S, Chemin K, Maciorowski Z, Lim A, Mazerolles F, et al. Hypomorphic mutation of ZAP70 in human results in a late onset immunodeficiency and no autoimmunity. Eur J Immunol. 2009; 39(7):1966–76. https://doi.org/10.1002/eji.200939385 PMID: 19548248

55. Schroeder ML, Triggs-Raine B, Zelinski T. Genotyping an immunodeficiency causing c.1624-11G>A ZAP70 mutation in Canadian Mennonites. BMC Med Genet. 2016; 17(1):50. https://doi.org/10.1186/s12881-016-0312-4 PMID: 27448562

56. Andreotti AH, Schwartzberg PL, Joseph RE, Berg LJ. T-cell signaling regulated by the Tec family kinase, Itk. Cold Spring Harb Perspect Biol. 2010; 2(7):a002287. https://doi.org/10.1101/cshperspect.a002287 PMID: 20519342

57. Huck K, Feyen O, Niehues T, Ruschendorf F, Hubner N, Laws HJ, et al. Girls homozygous for an IL-2-inducible T cell kinase mutation that leads to protein deficiency develop fatal EBV-associated lymphoproliferation. J Clin Invest. 2009; 119(5):1350–8. https://doi.org/10.1172/JCI37901 PMID: 19425169

58. Stepensky P, Weintraub M, Yanir A, Revel-Vilk S, Krux F, Huck K, et al. IL-2-inducible T-cell kinase deficiency: clinical presentation and therapeutic approach. Haematologica. 2011; 96(3):472–6. https://doi.org/10.3324/haematol.2010.033910 PMID: 21109689

59. Linka RM, Risse SL, Bienemann K, Werner M, Linka Y, Krux F, et al. Loss-of-function mutations within the IL-2 inducible kinase ITK in patients with EBV-associated lymphoproliferative diseases. Leukemia. 2012; 26(5):963–71. https://doi.org/10.1038/leu.2011.371 PMID: 22289921

60. Ghosh S, Bienemann K, Boztug K, Borkhardt A. Interleukin-2-inducible T-cell kinase (ITK) deficiency—clinical and molecular aspects. J Clin Immunol. 2014; 34(8):892–9. https://doi.org/10.1007/s10875-014-0110-8 PMID: 25339095

61. Schaeffer EM, Debnath J, Yap G, McVicar D, Liao XC, Littman DR, et al. Requirement for Tec kinases Rlk and Itk in T cell receptor signaling and immunity. Science. 1999; 284(5414):638–41. PMID: 10213685

62. Sahu N, Venegas AM, Jankovic D, Mitzner W, Gomez-Rodriguez J, Cannons JL, et al. Selective expression rather than specific function of Txk and Itk regulate Th1 and Th2 responses. J Immunol. 2008; 181(9):6125–31. PMID: 18941202

63. Kosaka Y, Felices M, Berg LJ. Itk and Th2 responses: action but no reaction. Trends Immunol. 2006; 27(10):453–60. https://doi.org/10.1016/j.it.2006.08.006 PMID: 16931156

64. Fowell DJ, Shinkai K, Liao XC, Beebe AM, Coffman RL, Littman DR, et al. Impaired NFATc translocation and failure of Th2 development in Itk-deficient CD4+ T cells. Immunity. 1999; 11(4):399–409. PMID: 10549622

65. Altman A, Villalba M. Protein kinase C-theta (PKC theta): a key enzyme in T cell life and death. Journal of Biochemistry. 2002; 132(6):841–6. PMID: 12473184

66. Stepensky P, Keller B, Buchta M, Kienzler A-K, Elpeleg O, Somech R, et al. Deficiency of caspase recruitment domain family, member 11 (CARD11), causes profound combined immunodeficiency in human subjects. J Allergy Clin Immunol. 2013; 131(2):477-+. https://doi.org/10.1016/j.jaci.2012.11.050 PMID: 23374270

67. Jabara HH, Ohsumi T, Chou J, Massaad MJ, Benson H, Megarbane A, et al. A homozygous mucosa-associated lymphoid tissue 1 (MALT1) mutation in a family with combined immunodeficiency. J Allergy Clin Immunol. 2013; 132(1):151–8. https://doi.org/10.1016/j.jaci.2013.04.047 PMID: 23727036

68. Torres JM, Martinez-Barricarte R, Garcia-Gomez S, Mazariegos MS, Itan Y, Boisson B, et al. Inherited BCL10 deficiency impairs hematopoietic and nonhematopoietic immunity. J Clin Invest. 2014; 124(12):5239–48. https://doi.org/10.1172/JCI77493 PMID: 25365219

69. Turvey SE, Durandy A, Fischer A, Fung SY, Geha RS, Gewies A, et al. The CARD11-BCL10-MALT1 (CBM) signalosome complex: Stepping into the limelight of human primary immunodeficiency. J Allergy Clin Immunol. 2014; 134(2):276–84. https://doi.org/10.1016/j.jaci.2014.06.015 PMID: 25087226

70. Hacker H, Karin M. Regulation and function of IKK and IKK-related kinases. Science's STKE: signal transduction knowledge environment. 2006; 2006(357):re13. https://doi.org/10.1126/stke.3572006re13 PMID: 17047224

71. Sun L, Deng L, Ea CK, Xia ZP, Chen ZJ. The TRAF6 ubiquitin ligase and TAK1 kinase mediate IKK activation by BCL10 and MALT1 in T lymphocytes. Mol Cell. 2004; 14(3):289–301. PMID: 15125833

72. Li Y, He X, Wang S, Shu HB, Liu Y. USP2a positively regulates TCR-induced NF-kappaB activation by bridging MALT1-TRAF6. Protein Cell. 2013; 4(1):62–70. https://doi.org/10.1007/s13238-012-2120-8 PMID: 23264041

73. Shinohara H, Kurosaki T. Comprehending the complex connection between PKC beta, TAK1, and IKK in BCR signaling. Immunol Rev. 2009; 232:300–18. https://doi.org/10.1111/j.1600-065X.2009.00836.x PMID: 19909372

74. Fusco F, Pescatore A, Conte MI, Mirabelli P, Paciolla M, Esposito E, et al. EDA-ID and IP, two faces of the same coin: how the same IKBKG/NEMO mutation affecting the NF-kappaB pathway can cause immunodeficiency and/or inflammation. Int Rev Immunol. 2015; 34(6):445–59. https://doi.org/10.3109/08830185.2015.1055331 PMID: 26269396

75. Johnston AM, Niemela J, Rosenzweig SD, Fried AJ, Delmonte OM, Fleisher TA, et al. A Novel Mutation in IKBKG/NEMO Leads to Ectodermal Dysplasia with Severe Immunodeficiency (EDA-ID). J Clin Immunol. 2016; 36(6):541–3. https://doi.org/10.1007/s10875-016-0309-y PMID: 27368913

76. Puel A, Reichenbach J, Bustamante J, Ku CL, Feinberg J, Doffinger R, et al. The NEMO mutation creating the most-upstream premature stop codon is hypomorphic because of a reinitiation of translation. Am J Hum Genet. 2006; 78(4):691–701. https://doi.org/10.1086/501532 PMID: 16532398

77. Pannicke U, Baumann B, Fuchs S, Henneke P, Rensing-Ehl A, Rizzi M, et al. Deficiency of innate and acquired immunity caused by an IKBKB mutation. N Engl J Med. 2013; 369(26):2504–14. https://doi.org/10.1056/NEJMoa1309199 PMID: 24369075

78. Mousallem T, Yang J, Urban TJ, Wang H, Adeli M, Parrott RE, et al. A nonsense mutation in IKBKB causes combined immunodeficiency. Blood. 2014; 124(13):2046–50. https://doi.org/10.1182/blood-2014-04-571265 PMID: 25139357

79. Nielsen C, Jakobsen MA, Larsen MJ, Muller AC, Hansen S, Lillevang ST, et al. Immunodeficiency associated with a nonsense mutation of IKBKB. J Clin Immunol. 2014; 34(8):916–21. https://doi.org/10.1007/s10875-014-0097-1 PMID: 25216719

80. Orange JS, Levy O, Brodeur SR, Krzewski K, Roy RM, Niemela JE, et al. Human nuclear factor kappa B essential modulator mutation can result in immunodeficiency without ectodermal dysplasia. J Allergy Clin Immunol. 2004; 114(3):650–6. https://doi.org/10.1016/j.jaci.2004.06.052 PMID: 15356572

81. Doffinger R, Smahi A, Bessia C, Geissmann F, Feinberg J, Durandy A, et al. X-linked anhidrotic ectodermal dysplasia with immunodeficiency is caused by impaired NF-kappaB signaling. Nat Genet. 2001; 27(3):277–85. https://doi.org/10.1038/85837 PMID: 11242109

82. Jorgensen SE, Bottger P, Kofod-Olsen E, Holm M, Mork N, Orntoft TF, et al. Ectodermal dysplasia with immunodeficiency caused by a branch-point mutation in IKBKG/NEMO. J Allergy Clin Immunol. 2016; 138(6):1706–9 e4. https://doi.org/10.1016/j.jaci.2016.05.030 PMID: 27477329

83. Kawai T, Nishikomori R, Izawa K, Murata Y, Tanaka N, Sakai H, et al. Frequent somatic mosaicism of NEMO in T cells of patients with X-linked anhidrotic ectodermal dysplasia with immunodeficiency. Blood. 2012; 119(23):5458–66. https://doi.org/10.1182/blood-2011-05-354167 PMID: 22517901

84. Chen ZJ. Ubiquitination in signaling to and activation of IKK. Immunol Rev. 2012; 246(1):95–106. https://doi.org/10.1111/j.1600-065X.2012.01108.x PMID: 22435549

85. Sun SC. The noncanonical NF-kappaB pathway. Immunol Rev. 2012; 246(1):125–40. https://doi.org/10.1111/j.1600-065X.2011.01088.x PMID: 22435551

86. Yamamoto M, Sato S, Saitoh T, Sakurai H, Uematsu S, Kawai T, et al. Pivotal function of Ubc13 in thymocyte TCR signaling. J Immunol. 2006; 177(11):7520–4. PMID: 17114420

87. Xiao G, Harhaj EW, Sun SC. NF-kappaB-inducing kinase regulates the processing of NF-kappaB2 p100. Mol Cell. 2001; 7(2):401–9. PMID: 11239468

88. Willmann KL, Klaver S, Dogu F, Santos-Valente E, Garncarz W, Bilic I, et al. Biallelic loss-of-function mutation in NIK causes a primary immunodeficiency with multifaceted aberrant lymphoid immunity. Nat Commun. 2014; 5:5360. https://doi.org/10.1038/ncomms6360 PMID: 25406581

89. Matsumoto M, Yamada T, Yoshinaga SK, Boone T, Horan T, Fujita S, et al. Essential role of NF-kappa B-inducing kinase in T cell activation through the TCR/CD3 pathway. J Immunol. 2002; 169(3):1151–8. PMID: 12133934

90. R-Core-Team. R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing; 2016. http://www.R-project.org.

91. Kohl M, Wiese S, Warscheid B. Cytoscape: Software for Visualization and Analysis of Biological Networks. Methods Mol Biol. 2011; 696:291–303. https://doi.org/10.1007/978-1-60761-987-1_18 PMID: 21063955

92. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006; Complex Systems:1695.

93. Boole G. The Calculus of Logic. Cambridge and Dublin Mathematical Journal. 1848; 3:183–98.

94. Thomas R. Boolean formalization of genetic control circuits. J Theor Biol. 1973; 42(3):563–85. PMID: 4588055

95. Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ. Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. BMC Syst Biol. 2009; 3:98. https://doi.org/10.1186/1752-0509-3-98 PMID: 19785753