CRANFIELD UNIVERSITY

Cranfield Health
Applied Bioinformatics

MSc

Academic Year 2010 - 2011

Marie C. Weston

A Systems Biology Approach To Target Discovery In Regulatory T Cells

Supervisor: Matt Page,
UCB Pharma, Slough, UK

Cranfield Supervisor: Michael Cauchi

August 2011

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science

ABSTRACT

Regulatory T cells (Tregs) have a central role in the maintenance of tolerance to self-antigens and the prevention of autoimmune disease. This study used an integrative systems biology approach to identify tolerogenic genes in Tregs which could potentially serve as novel therapeutic targets for immunological disorders.

A consensus Treg gene signature was generated by comparing gene expression in Treg vs naïve or conventional T cells across multiple public studies. Ingenuity Pathway Analysis software was then used to expand the Treg consensus gene list to include interacting proteins accessible to intervention by antibody therapeutics.

Many viruses co-opt genes for host proteins that modulate the host's immune system. It is hypothesized that some viruses may have co-opted genes that can induce tolerance, allowing the virus to evade elimination by the host's immune system. Putative tolerogenic genes were therefore selected for further investigation based upon their presence in viral genomes. The presence of human genes in viral genomes was investigated by performing a batch reciprocal BLAST search.

The biological significance of the human vs viral alignments was evaluated by manual inspection of the alignments and searching for the presence of shared motifs and protein family domains in the viral and human sequences.

A final list of ten putative tolerogenic genes included genes known to be associated with immune function and some already established therapeutic targets for autoimmune diseases, as well as four potentially novel therapeutic targets.

The biological rationale for the putative targets' involvement in tolerance was explored in the context of Treg gene expression and protein-protein interaction (PPI) network topology. A PPI network was generated and annotated with confidence scores for each of the interactions. The Cytoscape plugin JActiveModules was used to find putative functional network modules.

ACKNOWLEDGEMENTS

I would like to thank Matt Page for all his help and support throughout my time at UCB, and for giving me the opportunity to take on an interesting and stimulating project. Thanks also to Michael Palmowski for providing immunological guidance and for sharing some fascinating biological insights. Thank you also to my supervisor at Cranfield University, Michael Cauchi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

APC	Antigen Presenting Cells
BiNGO	Biological Networks Gene Ontology tool
BLAST	Basic Local Alignment Search Tool
cAMP	3'-5'-cyclic adenosine monophosphate
CCR7	Chemokine (C-C motif) Receptor 7
CD4	Cluster Of Differentiation 4
CD25	Cluster Of Differentiation 25
CD39	Cluster Of Differentiation 39 (also ectonucleoside triphosphate diphosphorylase 1)
CD73	Cluster Of Differentiation 73 (also ecto-5'-nucleotidase)
CD80	Cluster Of Differentiation (also B7-1)
CTLA-4	Cytotoxic T Lymphocyte-Associated Antigen 4
CXCR2	Chemokine (C-X-C motif) Receptor 2
DC	Dendritic Cell
DDI	Domain-Domain Interaction
DIP	Database of Interacting Proteins
EBI	European Bioinformatics Institute
EBV	Epstein Barr Virus
Foxp3	Forkhead Box p3
GEO	Gene Expression Omnibus
GO	Gene Ontology
GWAS	Genome Wide Association Studies
HCMV	Human Cytomegalovirus
HLA	Human Leukocyte Antigen
HPI	Human Proteome Initiative
HPR	Highest PMID re-use
HSP	High-Scoring Segment Pair
HUPO-PSI	Human Proteome Organization Proteomics Standards Initiative
IBD	Inflammatory Bowel Dsease

IPA	Ingenuity Pathway Analysis
IDO	Indoleamine 2,3-dioxygenase
IFNγ	Interferon gamma
IL-1	Interleukin 1
IL-1R	Interleukin 1 Receptor
IL-2	Interleukin 2
IL-4	Interleukin 4
IL-10	Interleukin 10
IL-35	Interleukin 35
IPEX	Immune dysregulation, Polyendocrinopathy, Enteropathy, X-Linked Syndrome
GITR	Glucocorticoid-Induced TNF Receptor Family-Related Gene
KEGG	Kyoto Encyclopedia of Genes and Genomes
LPR	Lowest PMID re-use
LRRC32	Leucine Rich Repeat Containing 32
MINT	Molecular Interactions Database
MIQL	Molecular Interaction Query Language
mTOR	Mammalian Target of Rapamycin
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
PMID	Pubmed ID
PPI	Protein-Protein Interaction
PSI	Proteomics Standards Initiative
PSI-MI XML	PSI (Proteomics Standards Initiative) Molecular Interaction XML
PSI-MI TAB	PSI (Proteomics Standards Initiative) Molecular Interaction Tabular
PSICQUIC	PSI (Proteomics Standards Initiative) Common Query Interface
PSISCORE	PSI (Proteomics Standards Initiative) Confidence Soring System
RA	Rheumatoid Arthritis
SMD	Stanford Microarray Database
SNP	Single Nucleotide Polymorphisms
TAP	Tandem Affinity Purification

TAP-MS	Tandem Affinity Purification- Mass Spectrometry
TCR	T cell Receptor
TGF-β	Transforming Growth Factor Beta
TNFα	Tumour Necrosis Factor Alpha
Treg cells	Regulatory T Cells
Th cells	T Helper Cells
TGOI-1	Tolerogenic Gene Of Interest-1 (anonymized gene in results)
TGOI-2	Tolerogenic Gene Of Interest -2 (anonymized gene in results)
TGOI-3	Tolerogenic Gene Of Interest -3 (anonymized gene in results)
TGOI-4	Tolerogenic Gene Of Interest -4 (anonymized gene in results)
TIGIT	T Cell Immunoreceptor with Ig and ITIM Domains
Y2H	Yeast Two Hybrid

1 Background and introduction

1.1 Regulatory T cells

1.1.1 Discovery and identification of regulatory T cells

The major role of the mammalian immune system is to protect the host from infection by pathogenic microorganisms. As with any reaction, there needs to be an opposing regulatory process in place to maintain homeostasis. Thus, the immuno-protective reactions must simultaneously be balanced by suppressive mechanisms to prevent inappropriate and excessive immune responses which could be harmful to the host. Regulatory T cells (Treg cells) have evolved as a subset of T lymphocytes whose function is to modulate potentially deleterious activities of conventional T helper (Th) or effector cells (Sakaguchi *et al.*, 2008).

Immunological tolerance is particularly important in the case of self-tolerance. The first line of self-tolerance is the elimination by negative selection of self-reactive T lymphocytes in the thymus, and B lymphocytes in the bone marrow. Regulatory T cells modulate potentially deleterious activities of conventional T helper cells that escape this central self-reactivity checking mechanism. Although the primary role of Treg cells is generally considered to be that of preventing autoimmune disease by maintaining self-tolerance, a number of additional immunosuppressant functions have been suggested as being attributable to Tregs for the prevention of immune system responsiveness under inappropriate conditions (Corthay, 2009). These include the suppression of responses against innocuous environmental substances seen in allergy and asthma, induction of oral tolerance to dietary antigens, induction of maternal tolerance to the foetus, protection of commensal bacteria from elimination by the immune system, suppression of pathogen-induced immunopathology, suppression of T cell activation triggered by weak stimuli, and feedback control of the magnitude of the immune response by effector T cells (Corthay, 2009).

The concept of T cell-mediated suppression of immune responses has been around for over forty years. In 1969, Nishizuka and Sakakura demonstrated that thymectomy of normal mice on the third day of life resulted in the development of organ-specific autoimmune disease, resulting in destruction of the ovaries. Thymectomy on days 1 or 7 of life did not result in autoimmune disease. This led to the hypothesis that during the first 3 days of life, autoreactive T cells are exported from the thymus; but that later, between days 3 and 7, a subset of immunosuppressive T cells emigrate from the thymus into the circulation which modulate the autoreactive T cells. If the day 3 mice later received transplanted thymocytes or splenocytes from normal mice, the disease was prevented.

In a seminal 1970 study, Gershon and Kondo demonstrated the importance of antigen-specific suppressor T cells. However, the subsequent failure to determine specific markers to distinguish these suppressor T cells from other T cells, the inability to purify them, and uncertainty over the molecular mechanisms of their immunosuppressive activity raised doubts about the existence of a distinct lineage, and as a consequence, interest in these suppressor T cells waned. Much later, in 1995, Sakaguchi *et al.* separated a fraction of CD4+ cells which constituted 5-10% of peripheral CD4+ cells and constitutively expressed high levels of CD25 (IL2 receptor α chain). They demonstrated that this fraction of cells could prevent autoimmune disease in mice. Depletion of these $CD4^+CD25^+$ cells in mice resulted in autoimmune diseases in multiple organs, and subsequent transfer of $CD4^+CD25^+$ cells prevented the autoimmunity (Figure 1.1). A later study (Asano *et al.*, 1996) revealed that these $CD4^+CD25^+$ cells did not appear in the periphery until 3 days after birth, and that they could prevent autoimmunity in 3 day thymectomized mice when given one week after the thymectomy, correlating nicely with the 1969 study by Nishizuka and Sakakura. Thus, a specific subset of T cells was finally defined which were eventually named regulatory T cells. A further six years elapsed before the confirmation of their existence in humans (Shevach, 2001). It has since been found that the identification of human Treg cells is more problematic, as CD25 is expressed on around a quarter of $CD4^+$ T cells, and it is thought that only the very highest CD25 expressers possess significant suppressive properties (Shevach, 2006). Hence, simple $CD4^+CD25^+$ sorting is not sufficient to precisely identify human Treg populations.

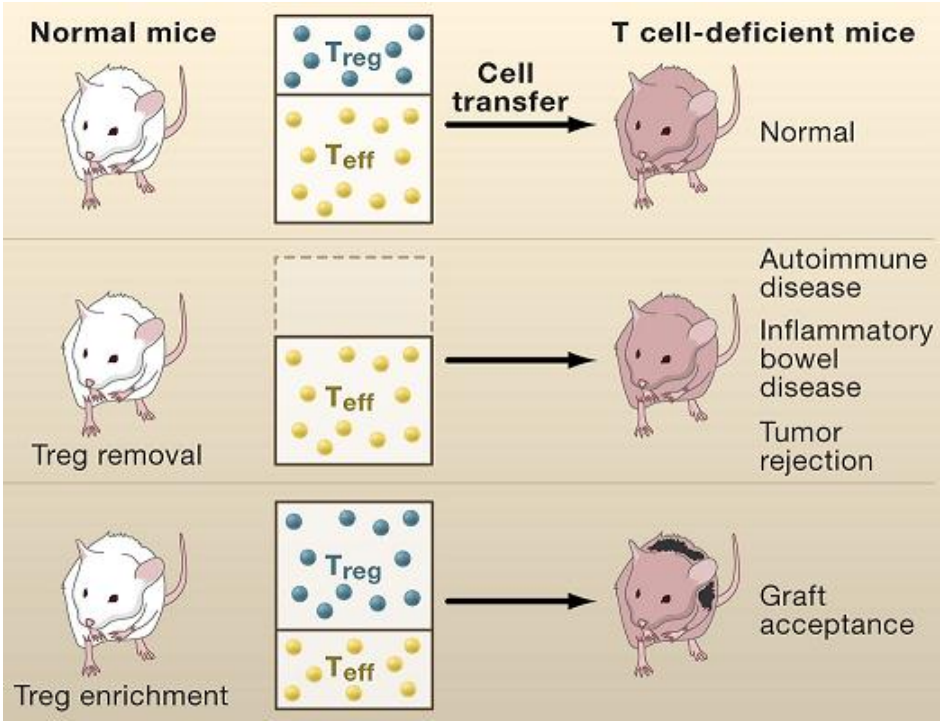


Figure 1.1 Effects of Treg deficiency in mice. T cell suspensions taken from normal mice can be depleted of $CD25^{+}CD4^{+}$ Treg cells and transferred to syngeneic T-cell deficient mice (such as athymic nude mice). The recipient mice spontaneously develop autoimmune disease and inflammatory bowel disease and reject tumour cells. When $CD25^{+}CD4^{+}$ Tregs are enriched from normal mice and transferred, the recipient mice, in addition to inhibition of autoimmune disease, will accept allogeneic skin grafts.

Figure from Sakaguchi et al. (2008)

A further significant landmark in Treg research was the discovery of Foxp3 (forkhead box p3), a member of the forkhead/winged helix family of transcription factors. Foxp3 is expressed in naturally occurring Treg cells and is the major regulator of Treg cell development and function. In 2001, the Foxp3 gene was identified as the defective gene in the Scurfy mouse, a mutant mouse strain which spontaneously develops severe autoimmune disease and inflammation (Brunkow *et al.*, 2001). The autoimmunity and inflammation was caused by hyperactivity of CD4+ cells and over-production of pro-inflammatory cytokines. Mutations of the human orthologue, FOXP3, cause a similar disease, IPEX (immune dysregulation, polyendocrinopathy, enteropathy, X-linked

syndrome), which results in multiorgan autoimmune disease, severe allergy and inflammatory bowel disease (Sakaguchi *et al.*, 2007). Similarities between the autoimmunity and inflammation produced by Foxp3 defects, and disease caused by depleting or manipulating $CD25^+CD4^+$ Treg cells prompted the investigation into the possible contribution of Foxp3 to the development or function of Treg cells (Hori *et al.*, 2003). It was observed that expression of Foxp3 mRNA was predominantly restricted to the $CD25^+CD4^+$ population of cells in both the thymus and the periphery. *In vitro* stimulation of $CD25^-CD4^+$ cells failed to elicit Foxp3 expression, but retroviral gene transfer of Foxp3 into $CD25^-CD4^+$ into immature T cells converted them into Treg-like cells which displayed similar behaviour to Tregs. This study indicated that Foxp3 expression did not occur as a consequence of cell activation, but was necessary for the Treg cell development. Foxp3 has been referred to as a master controller of the development and function of natural $CD25^+CD4^+$ Treg cells (Sakaguchi *et al.*, 2007), but this hypothesis has been challenged (Hori, 2008). It has been suggested that rather than being the master controller of Treg cell lineage, the function of Foxp3 may be to amplify and fix pre-established Treg-associated molecular features (Gavin et al, 2007)

In contrast to mice, Foxp3 expression in humans may not be completely confined to $CD25^+CD4^+$ cells. It has been observed that human $CD25^-CD4^+$ T cells may express Foxp3 mRNA upon T cell receptor (TCR) stimulation, although the expression levels are generally much lower and more transient than in natural Treg cells (Walker *et al.*, 2003; Morgan *et al.*, 2005). This raises some issues around the notion of Foxp3 being used as a definitive Treg marker for human cells.

Foxp3 appears to control Treg cells by activating or repressing the expression of many genes, either directly or indirectly, by binding to other transcription factors such as NF-AT, NF-kB and AML/Runx1 (Carson *et al.*, 2006). For example, Foxp3 binding may inhibit IL-2 (interleukin 2), IFNγ (interferon gamma) and IL-4 (interleukin 4) expression, and increase expression of cell surface molecules such as CD25, CTLA-4 (cytotoxic T lymphocyte-associated antigen 4) and GITR (glucocorticoid-induced TNF receptor family-related gene), all phenotypes associated with the Treg cells (Carson *et al.*, 2006).

There are now thought to be many subpopulations of suppressive T cells, including IL-10-producing suppressor cells Tr1 cells, TGF-β-producing Th3 cells, and $CD8^+$ T cells (Shevach, 2006, Tang & Bluestone, 2008). In addition, naïve T cells from the periphery can be induced to express Foxp3 by *in vitro* T cell receptor stimulation in the presence of IL-2 and TGF-β. In mice, these adapt Treg-like immunosuppressive behaviour (Huter *et al.*, 2008), whereas in humans, these $CD4^+Foxp3^+$ cells induced in this manner do not possess suppressive properties (Tran *et al.*, 2007). Non-induced centrally-derived $CD25^+CD4^+Foxp3^+$ Treg cells develop in the thymus and are crucial in the maintenance of immune homeostasis and self-tolerance. These are also referred to as natural Tregs, and are the cells upon which this project thesis will be focussed. Therefore this introduction will concentrate on this cell type.

1.1.2 Mechanisms of immune suppression by regulatory T cells

The molecular mechanisms by which Treg cells suppress the actions of other immune cells have still not been fully elucidated, but multiple modes of suppression have been proposed (Sakauchi *et al.*, 2009). Tregs are able to repress the proliferation of antigen-stimulated naïve T cells and prevent their differentiation to effector T cells, as well as suppressing the effector activities of differentiated $CD4^+$ and $CD8^+$ T lymphocytes. In addition, Tregs can repress the activities of multiple immune cells, including B lymphocytes, macrophages, dendritic cells, natural killer cells, natural killer T cells, and osteoclasts (Sakaguchi *et al.*, 2009). It is uncertain whether there is a common core suppressive mechanism shared by all Treg cells at any time or location, whether particular mechanisms kick into action dependent upon the situation or conditions of the immune response, or whether these different mechanisms serve to act together, synergistically. Suggested mechanisms of immunosuppression by Tregs include:

(i) Secretion of immunosuppressive cytokines: Treg cells produce immunosuppressive cytokines such as IL-10, TGF-β, galectin-1 and IL-35 (Tang & Bluestone, 2008). It has been shown that in mice with inflammatory bowel disease (IBD) induced by

depletion of Treg cells, IL-10 and TGF-β contribute to the suppression of the disease (Read *et al.*, 2000; Sakaguchi *et al.*, 2009). In mouse *in vivo* models, neutralization of TGF-β and antagonism of the IL-10 receptor prevents Treg-mediated suppression of IBD, type I diabetes, leishmania skin infection and transplantation (Tang & Bluestone, 2008).

It has also been proposed that Treg cells can induce cytokine deprivation-induced apoptosis of effector T cells by 'consumption' of IL-2. Tregs express high levels of CD25, the high affinity receptor for IL-2, so they may act as an IL-2 'sink' by competing with effector T cells for IL-2 (Pandiyan *et al.*, 2007).

(ii) Cell to cell contact with effector T cells: *In vitro* experiments demonstrating the ability of Treg cells to suppress effector T cell proliferation in the absence of antigen-presenting cells [APCs] (Picirillo & Shevach, 2001), and others that showed their inability to suppress proliferation when the two populations are separated by a semi-permeable membrane (Thornton & Shevach, 1998), led to the proposal that Treg cells suppress through direct contact with effector T cells. One suggested mechanism for suppression through cell contact is the killing of effector cells directly through the release of granzyme B and perforin (Grossman *et al.*, 2004; Gondek *et al.*, 2005). Other modes of suppression include mechanisms mediated through the release of negative signals, such as cyclic AMP (cAMP) and adenosine. cAMP is a potently immunosuppressive signalling molecule which has been shown to be present in large amounts in the cytoplasm of Treg cells. It has been suggested that cAMP could be passed onto effector T cells by contact through gap junctions (Bopp *et al.*, 2007). Adenosine can bind to adenosine receptors on effector T cells to suppress their activity. CD39 (ectonucleoside triphosphate diphosphorylase 1) and CD73 (ecto-5'-nucleotidase) which are expressed on the surface of Tregs catalyse the generation of extracellular adenosine (Deaglio *et al.*, 2007). However some studies utilizing imaging analysis have demonstrated that Tregs and effector T cells do not interact stably during *in vitro* and *in vivo* suppression (Tang *et al.*, 2006, Tang & Krummel, 2006).

(iii) Modulation of antigen presenting cells: Tregs can directly interact with dendritic cells (DCs). Aggregation of Tregs around DCs allows them to out-compete antigen-

specific effector T cells, affecting the ability of effector T cells to engage and become activated by the DCs (Sakaguchi *et al.*, 2008). Activated Tregs interacting with the DC may also inhibit DC functions such as antigen presentation. The imaging studies by Tang *et al.* (2006) mentioned in the previous paragraph, while showing that Tregs and effector T cells do not interact with each other stably, did demonstrate that Tregs and DCs interact, and that their interactions prevent the formation of stable conjugates between effector T cells and DCs. Tregs may also stimulate DCs to produce immunosuppressive factors. Tregs can stimulate DCs to increase the production of the enzyme, IDO (indoleamine 2,3-dioxygenase), which catalzyes the conversion of tryptophan to kynurenine, which is toxic to neighbouring T cells (Puccetti & Grohmann, 2007; Tang & Bluestone, 2008). Tregs may downregulate the expression of CD80 and CD86 on APCs, which are co-stimulatory molecules that provide signals required for priming and activation of T cells (Sakaguchi *et al.*, 2008). The production of IDO and the downregulation of CD80/86 expression are both dependent on CTLA-4 (Sakaguchi *et al.*, 2009).

1.1.3 Regulatory T cells as therapeutic targets

The preceding sections have illustrated how depletion of Treg cells can lead to a variety of autoimmune diseases, as well as inflammatory and allergic diseases. It can thus be inferred that autoimmune disease may develop as a result of Treg cell dysfunction or alterations in the balance between Tregs and self-reactive T cells. Treg cell depletion can also have some potentially beneficial effects for the host, such as enhanced immunity to tumours and enhanced antimicrobial immunity. Treg cells may therefore be exploited as therapeutic targets for the modulation of pathological immune responses. Augmentation of the functions, development or survival of Tregs may be useful for immunosuppressive therapy of diseases such as autoimmunity, allergy and IBD, and also the enhancement of foetal-maternal tolerance and the establishment of tolerance to organ transplants. Conversely, reductions in Treg cell numbers, functions, their migration or survival could be of benefit for the treatment of cancers and chronic infection.

This project and thesis will focus on the identification of tolerogenic genes in Treg cells, so this section will discuss the potential targeting of Tregs to induce immunosuppression in autoimmune disorders. Currently, the most common approach to treating immunological diseases is the administration of immunosuppressive drugs. A number of pharmacological agents have been associated with the boosting of Treg activity (Ohkura *et al.*, 2011). Rapamycin is an immunosuppressive drug already in clinical use for the prevention of organ transplant rejection. It acts through inhibition of mTOR (mammalian target of rapamycin) signalling. *In vitro* studies have shown that rapamycin promotes expansion of Treg cells isolated from both healthy subjects and patients with type I diabetes, through the selective inhibition of proliferation of effector T cells (Battaglia *et al.*, 2006). The sphingosine 1-phosphate receptor agonist, FTY720 has also been used to prevent organ transplant rejection, and amongst its immunosuppressive mechanisms is an ability to increase Treg cell activity. Retinoic acid and aromatase inhibitors have been shown to respectively promote differentiation of naïve T cells to Tregs and expansion of Tregs (Ohkura *et al.*, 2011).

However, the use of immunosuppressant drugs often does not distinguish between beneficial and deleterious immune responses, as there is no antigen specificity. Immunosuppressive therapy is by necessity, continued throughout a patient's lifetime, as withdrawal of therapy may result in disease relapse. Pre-clinical experiments in mouse models have indicated that the adoptive transfer of Tregs may prevent or cure many immunological diseases by restoring self-tolerance (Roncarlo & Battaglia, 2007). The potential use of this type of therapy may have several advantages over conventional immunosuppressive drugs, as there would be antigen specificity without the general suppressive effects. Long term physiological regulation may be induced by the re-establishment of immunological homeostasis. The therapy would be patient-specific and could be tailored to the particular needs of the patient.

There are, however, a number of technical issues that would need to be considered with this type of therapy. The Tregs would need to be collected from blood and immediately processed. They would need to be purified to a single population. As the Foxp3+ natural Tregs represent only 5-10% of $CD4^+$ cells, they would need to be further expanded *in vitro*. Tregs are readily expandable in short term culture, but Foxp3

expression is not particularly stable in long term cultures (Miyara & Sakaguchi, 2011). These technical cell manipulations and reinfusion of the cells into the patient needs to be performed with good quality controls in place and under clinical GMP (good manufacturing practice) conditions, therefore appropriate facilities need to be in place, restricting the number of places equipped to perform this type of therapy (Roncarlo & Battaglia, 2007). As a result, this is likely to be an expensive, and therefore commercially unattractive therapeutic approach. Safety also has to be monitored. The possible conversion of $Foxp3^+$ Tregs into effector T cells, resulting in worsening of disease is a concern. It has been shown in mice that in an inflammatory environment, Tregs may lose Foxp3 expression and become pathogenic Th17 cells (Xu *et al.*, 2007, Miyara & Sakaguchi, 2011). The possibility of uncontrolled cell proliferation and pan-immunosuppression would also need to be monitored (Roncarlo & Battaglia, 2007). Some clinical trials in patients undergoing bone marrow therapy (Roncarlo & Battaglia, 2007) and human leukocyte antigen (HLA)-haploidentical hematopoietic stem-cell transplantation (Di Ianni *et al.*, 2011) have shown promising preliminary results, indicating the feasibility of this therapeutic strategy.

Roncarlo and Battaglia (2007) have suggested that in the future, it may be possible to induce or enhance Treg cell function or confer regulatory activity to effector T cells by *ex vivo* transfer of genes. The transfer of Foxp3 into non regulatory T cells, for example may be a possible candidate for consideration.

In conclusion, the central role of regulatory T cells in maintenance of self-tolerance and prevention of autoimmune disease make this cell type an attractive target for immunosuppressive drugs. It would appear that current therapeutic strategies for immunosuppression are either non-specific, or expensive and technically challenging. Therefore, there is a real need for novel therapeutics that would be more selective and consequently safer, as well as being cheap and easy to administer. The identification of potential tolerogenic genes and as a corollary, potentially novel molecular targets in regulatory T cells, can make a valuable initial contribution towards this goal.

1.2 Systems and network biology

1.2.1 Introduction to systems biology

Systems biology takes a holistic approach to understanding relationships between the components that make up the 'system', e.g. the organelle, cell, organ, or whole organism. The behaviours and functions that arise as a result of the interactions of the component parts, rather than as a result of a single part of the system, are referred to as the emergent properties of the system (Alberghina *et al.*, 2009). Traditionally, particularly in the areas of cellular and molecular biology, research has been carried out in an essentially reductionist manner, i.e. by examination of the basic molecular components of the system. Systems biology adopts a top-down approach to understanding the structural and dynamic organization of elements of a system at various levels of resolution, and how their interactions produce the emergent properties of the system.

The growth of systems biology over the last decade has been accelerated by a number of advances (Arrell & Terzic, 2010):

(i) The sequencing of the human genome in 2001.

(ii) The availability of comprehensive biological data repositories, e.g. genes, proteins, metabolites, protein interaction and biochemical pathway knowledge bases.

(iii) The development of robust high-throughput techniques, enabling large-scale detection, identification and assessment of molecular variability, and the consequent expansion in data generation. These large data sets have resulted from advances in technologies such as transcriptomics, proteomics, and metabolomics, and also advances in high-throughput imaging.

(iv) The development of new technologies and computational approaches to analyse large amounts of data, and to integrate with network data.

(v) The standardization of data formats and ontologies

(vi) The internet revolution, which has provided the means of rapid dissemination and acquisition of knowledge and data.

People generally understand two things by systems biology:

- The classical idea of systems biology which is quantitative network modelling and simulation
- Integrative approaches, known as Integrative Systems biology. This approach aims to maximally exploit broad-scale omic datasets, typically by using comprehensive protein-interaction networks as an integrative framework for the integration and computation of omic data.

1.2.2 Interaction networks

The development of high-throughput, quantitative, massively parallel technologies has provided collections of data which have delivered inventories of the cellular parts and information on the stoichiometries of the molecular components of the biological system. This has provided unparalleled resolution into the wiring of cellular signalling. Advances in molecular techniques has changed the organizational view of the cell from being a "bag of enzymes" to a network of complex, highly inter-connected molecular interactions (Vidal *et al.*, 2011). The full complement of these macromolecular interactions in cells constitutes what is known as the *interactome*. The challenge now is to assemble these components in a systematic manner into functional molecular networks that can be used to reveal information on fundamental biological processes, how they can become dysregulated by disease and thus predict how a network may respond to modulation by therapeutic intervention (Pe'er & Hacohen, 2011).

Biological network analysis simplifies these complex molecular interaction systems by representing them as a simplified mathematical object called a graph, comprising a collection of nodes and edges (Albert, 2005). Elements of the system such as proteins, genes or metabolites are represented as the nodes, and the relationships between them are represented by the edges, i.e. lines connecting them (Figure 1.2). The relationships may involve physical, regulatory or genetic relationships, the flow of material from a

substrate to a product, or other types of relationship, depending upon the type of network being represented. The edges may be uni-directional, bi-directional or non-directed. A non-directed edge for example, may be used to represent mutual interactions such as a direct physical PPI. Additional information may be assigned to the edges such as positive or negative signs representing activation or inhibition, respectively, reaction rates, or confidence levels; or different classes of nodes may be represented (Albert, 2005). A number of approaches can be taken to construct these interaction networks (Arrel & Terzic, 2010; Vidal *et al.*, 2011). For example:

- *de novo* from experimental results showing PPIs, using techniques such as yeast two hybrid (Y2H) screening, immunoprecipitation or tandem affinity purification (TAP)
- Compilation or curation of pre-existing data available in the literature
- Systematic high-throughput experimental mapping strategies; i.e. use of software tools that leverage data generated by the methods in the previous two points. Known interactions can be applied to an 'omic' data set using pathway analysis software such as Ingenuity Pathway Analysis (IPA), Cytoscape, MetaCore or Pathway Studio
- Computational predictions based upon information other than PPIs, such as sequence similarities, protein structures and co-regulated genes (e.g. based on correlated expression clusters predicted from gene expression data)
- Reverse engineering; i.e. the use of experimental data sets using numerous perturbations for reconstruction of the underlying structure of a network or the regulatory relationships between the nodes.

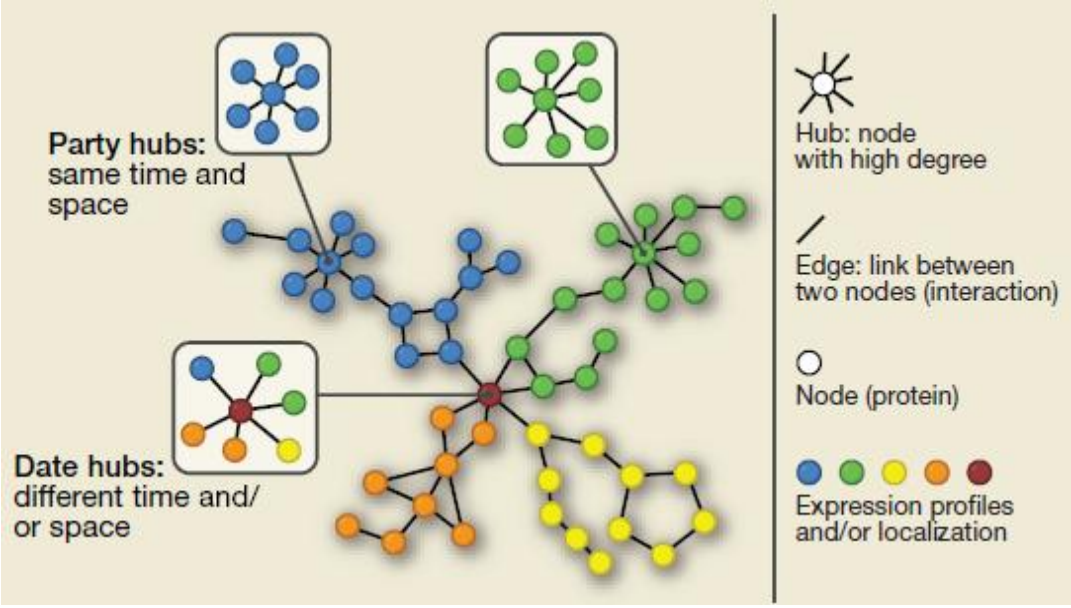


Figure 1.2 Network components. *Figure from Seebacher & Gavin (2011.)*

Interaction networks provide a global scaffold or template for the integration of multiple data-types and for overlaying qualitative or quantitative information, facilitating mathematical modelling and providing foci for hypothesis generation (Arrel & Terzic, 2010).

Network topology can provide valuable information for the identification of molecules that are functionally important or critical for network integrity and function. Most nodes have few connections, but some are highly connected to other nodes. These are called hubs. Proteins may vary their connections with time and location. There are two types of hub: "party hubs" and "date hubs" (Han *et al.*, 2004) [Figure 1.2]. A party hub co-expresses and/or co-localizes simultaneously with all its interacting nodes at the same time and in the same spatial location. Date hubs vary their connections with their interacting partners at different times and locations (Han *et al.*, 2004). Party hubs are more likely to be the module organizers, acting within the same biological process, whereas date hubs are more likely to be module connectors, linking biological processes. Information on these key nodes, particularly in disease-related networks could be useful in drug discovery. If a potential therapeutic target turns out to be a highly connected hub molecule, then it may be unsuitable as a drug target, as

modulation of its activity may affect many other activities within the cell, giving rise to unwanted off-target effects (Kann, 2007). Molecules corresponding to nodes with fewer connections which affect only the pathway involved in the disease may represent more suitable therapeutic targets.

1.2.3 Use of data mining and systems biology for target identification

The identification of suitable targets is crucial to the drug discovery process. Target selection without a firm mechanistic rationale is a leading cause of drug attrition rates because of poor pharmacokinetic profiles, unexpected toxicity, or because the biological hypothesis underlying the target and drug selection was flawed (Butcher, 2003). Therefore, during the process of identifying and selecting reliable druggable targets, it is important to gain insights into the underlying mechanisms and molecular interactions involved in disease processes, so that better targets can be selected. The construction of biological networks and predictive models, and the interpretation of quantitative 'omic' data in the biological context of PPIs can facilitate this process. This involves the gathering and integration of heterogeneous data types. The recent advances in systems biology have provided a new approach to target identification (Yang *et al.*, 2009). Current data mining approaches include text mining of literature databases, microarray, proteomic and chemogenomic data mining (Yang *et al.*, 2009). In the search for new targets in regulatory T cells during the present study, transcriptomic data from publicly available microarray studies will be used. Therefore, this section will be confined to the discussion of microarray data mining techniques.

Microarray data mining involves the application of bioinformatics to the analysis of microarray data for the identification of gene signatures and/or biological pathways that can define a particular phenotype, such as a disease (Ricke *et al.*, 2006). Two basic approaches that can be applied to gene expression data mining are unsupervised clustering and supervised classification. Unsupervised clustering is an exploratory approach which aims to determine whether groups of genes or biological samples share similar expression patterns. Supervised classification aims to identify genes that can

distinguish between known samples, phenotypes or classes, such as diseased vs non-diseased samples (Mount & Pandey, 2005).

The exponential growth in microarray data and the emergence of "open biology" has led to the growth in publicly available microarray datasets in open data repositories, such as NCBI Gene Expression Omnibus (GEO), Stanford Microarray Database (SMD) and EBI Array Express (Galperin & Cochrane, 2011). The availability of these collections of vast numbers of data points has made possible the meta-analysis of multiple transcriptomic datasets that address similar biological questions (Kupershmidt *et al.*, 2010). Many significant discoveries have been made through meta-analysis of multiple gene expression data sets. For example, the identification of consistently deregulated genes in prostate cancer (Rhodes *et al.*, 2002), the identification of gene expression profiles in breast cancer (Wirapati *et al.*, 2008) and the identification of candidate biomarkers for colorectal cancer (Chan *et al.*, 2008).

Individual data mining approaches used in isolation are not usually sufficient for constructing biological networks and delineating cellular processes. Improvements can be achieved by integration with other data sources, such as PPI data, localization information, published literature and phylogenetic information. It has been shown that integration of heterogeneous data types increases the accuracy of gene function predictions when compared with single high through-put methods used alone (Troyanskaya, 2005). For example, in the case of gene expression data, key regulators of a biological function may not be differentially expressed. Integration with interaction network information allows molecules interacting or connected within the same network as differentially expressed gene products to be implicated in a biological process (Figure 1.3). This principle is known as 'guilt by association'. Manually curated knowledge bases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and interactome databases such as MINT (Molecular Interactions database), DIP (Database of Interacting Proteins), BioGrid and InAct have allowed the integration and overlaying of gene expression data with known interactions, regulatory relationships and biochemical reactions, and their analysis and visualization in the context of biological networks. Specialized pathway analysis software such as Ingenuity Pathway

Analysis (IPA), Cytoscape, MetaCore or Pathway Studio has been developed to facilitate the mapping of gene expression data to biological networks.

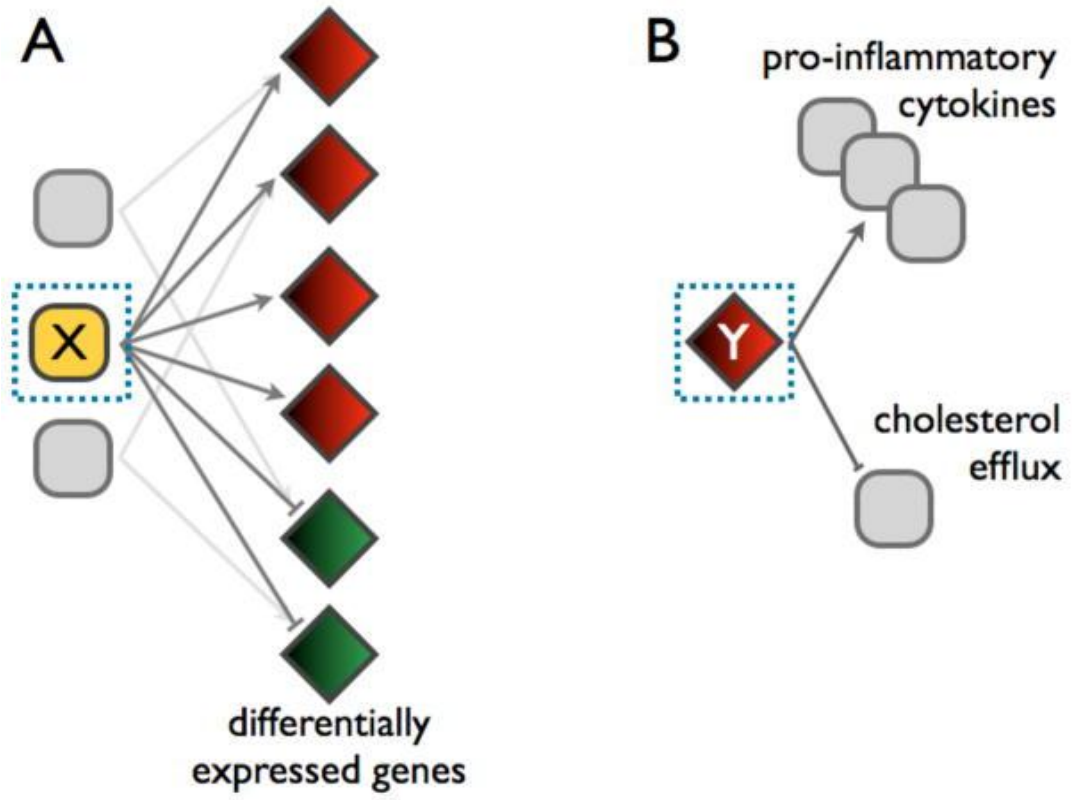


Figure 1.3 Regulatory molecules implicated by network analysis

A. A regulatory molecule "X" that is not differentially expressed can be implicated by its connectivity to differentially expressed molecules, in the interaction network.

B. A differentially expressed molecule "Y" can be identified as a key regulator through its connections to key disease-associated molecules and biological processes. Red/green colors denote up/downregulation of mRNA.

Figure from Ramsey et al. (2010)

2 Aims and objectives

The aim of this project is to identify tolerogenic genes in regulatory T cells which could potentially serve as novel therapeutic targets for immunological disorders. This will be achieved by utilizing an integrative systems biology approach which will combine gene expression and protein interaction data. Putative tolerogenic genes will ultimately be selected for further investigation based upon their presence in viral genomes, as viruses are likely candidates to have co-opted genes for host proteins that modulate the immune system of the host. It is hypothesized that some viruses, particularly those that reside in the host for a long period of time, may have co-opted genes that can induce tolerance, allowing the virus to evade elimination by the host's immune system. The identification of genes that may control tolerance mechanisms could potentially represent targets for a new class of therapeutic that could actually induce tolerance and provide long term or permanent control of immune disorders, or even the holy grail of long term remission.

These aims will be achieved using the following methods (Figure 2.1):

1) Generate a consensus human Treg cell gene signature

    - Gene expression data will be mined by selecting transcriptomic datasets available in public data repositories which compare gene expression in regulatory T cells and non-regulatory $CD4^+$ T cells. Meta-analysis across the datasets will be performed to generate a Treg consensus gene signature

2) Expand the consensus Treg gene list to include upstream interacting plasma membrane-resident proteins and proteins located in the extracellular space

- The Ingenuity Pathway Analysis software tool which contains a highly comprehensive, manually curated protein interaction network will be used for this network expansion.

3) Identify genes in the expanded list that have been integrated into viral genomes.

   - Download all available viral genome sequences (viruses hosted in humans) from NCBI and write a Perl script to automate the reciprocal blasting of the gene list against them.
   - Evaluate the human/viral sequence alignments for potential biological significance

Finally, the biological rationale for each putative target's involvement in tolerance will be explored within the context of the Treg gene expression data and the interaction network topology. This will be achieved by applying the following methods:

1) Generate a Treg PPI network with associated confidence scores. Protein interaction data will be obtained from public interaction data repositories by querying with the Treg gene signature, and interactions scored using the PSISCORE application.

2) Overlay gene expression data from the consensus Treg gene signature.

3) Identify active subnetworks or putative functional modules. Functional modules can be identified as highly connected network regions which show significant changes in gene expression.
   The Cytoscape plugin, JActiveModules was used, which utilizes algorithms to search for connections between gene expression and network topology.

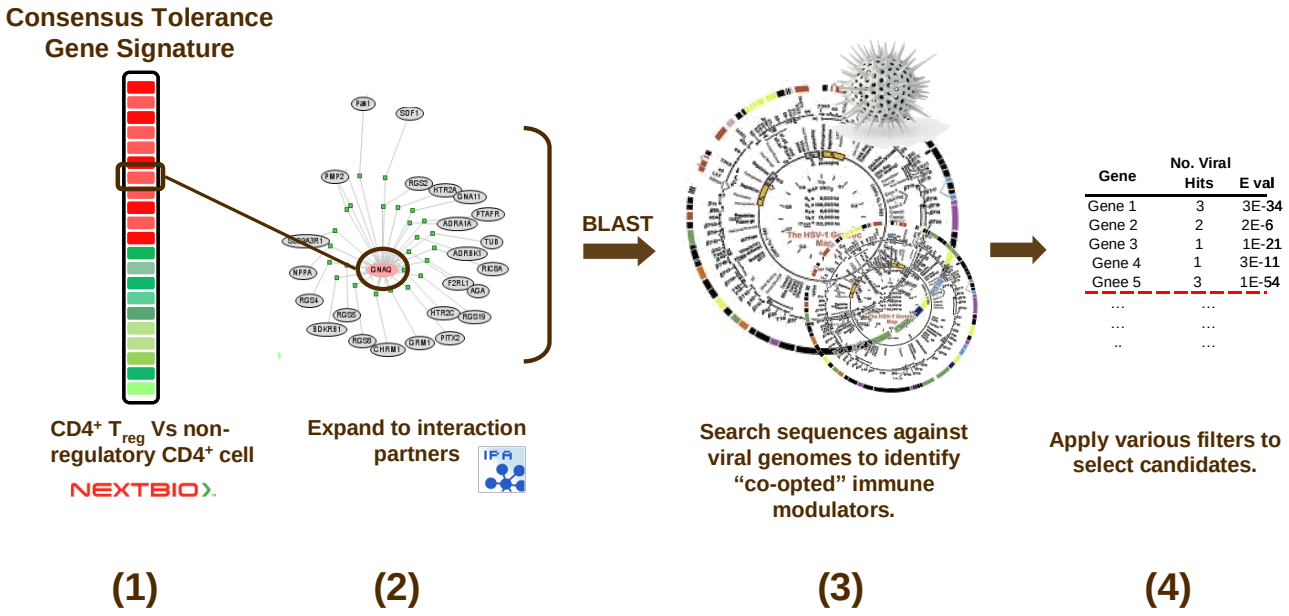


Figure 2.1 Workflow to identify tolerogenic genes

3 Methods

3.1 Generation of a consensus Treg gene signature

Gene expression data was mined by selecting transcriptomic datasets available in public repositories. The NextBio software tool was utilized (Kuperschmidt *et al.*, 2010). This is a data mining framework, containing a library of gene expression data sourced from several public databases, including Gene Expression Omnibus (GEO), Array Express, and Stanford Microarray database.

Studies were selected where Treg $CD4^+$ cells were compared with non-Treg (naïve or conventional) $CD4^+$ cells. Treg cells are CD4 and CD25-positive. However, around a quarter of all $CD4^+$ T cells express CD25, so in isolation, this is not a reliable Treg cell marker. Studies were therefore selected for inclusion in the analysis based upon the expression of certain key Treg-associated genes (genes selected in consultation with expert immunologists at UCB). The strict criteria for inclusion in the analysis were that Treg cells had to display increased expression of this particular combination of genes. Individually these genes may not be reliable markers for Treg cells, but together they are good indicators of a Treg phenotype.

Comparison of multiple studies investigating the same phenomenon allows the identification of consistently differentially expressed genes, a so-called consensus Treg gene signature; and mitigates the confounding effects of noise in individual gene expression studies. Meta-analysis across the selected Treg v non Treg datasets was performed to generate the Treg consensus gene signature. This allows the identification of the most consistently and highly regulated genes across multiple datasets. NextBio uses a combination of rank-based enrichment algorithms, ontologies and meta-analysis techniques to computationally identify gene signatures. The two most important parameters are the activity level of a gene in each dataset and the specificity (the number of datasets in which the gene is active). Genes which were differentially expressed (up or down) in Treg compared to non-Treg cells, in 4 of the 6 studies, were selected for subsequent analysis. This selects for genes that are differentially expressed

in a majority of the studies, but provides some degree of margin for allowances to be made for hybridization errors or other errors in individual studies.

3.2 Expansion of the Treg gene signature to include interacting proteins

Putative therapeutic targets may potentially be differentially expressed Treg proteins, or proteins interacting with a Treg protein that can modulate its activity. UCB has good therapeutic antibody generation capabilities, therefore putative target candidates were confined to proteins localized in either the plasma membrane or extracellular space. This makes them accessible for modulation by antibody therapeutics.

Identification of interacting proteins and the determination of their cellular localizations were carried out through the use of the Ingenuity Pathway Analysis tool [IPA] (Ingenuity Systems, USA, www.ingenuity.com). The dataset containing the consensus differentially expressed Treg gene identifiers and corresponding expression values was uploaded into the application. Each identifier was mapped to its corresponding object in the Ingenuity® Knowledge Base.

Proteins localized in the plasma membrane or extracellular space which directly interact with differentially expressed Treg protein products were identified. IPA provides information on cellular localizations of molecules, based upon Gene Ontology (GO) cellular compartment annotations. The IPA application allows a number of filters to be set when growing/building a network. A specific set of filters were applied to limit the output to molecules only fulfilling appropriate criteria.

Differentially expressed Treg proteins and their direct interactors which were expressed only in the plasma membrane, extracellular space, or an unknown location were selected. A list of these proteins and corresponding Entrez gene IDs was exported from IPA.

3.3 Identification of proteins showing homology with viral proteins

3.3.1 Standalone BLAST

Differentially expressed Treg genes and their interacting partners were filtered based upon their presence in virus genomes. Proteins showing homology with viral proteins were identified by performing BLAST (Basic Local Alignment Search Tool) searches against a database of viral proteins. Highest scoring viral hits, i.e. the viral protein showing the greatest homology to the human protein were re-blasted against a database of human proteins to confirm the hits. This reciprocal BLAST was to ensure that the hits were real and that the highest scoring alignment was not, for example, just an alignment with another member of the same protein family which is not the true homologue.

To this end, a standalone NCBI BLAST application was used. This can be freely downloaded from the NCBI website:

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/

3.3.2 Creation of local viral and human BLAST databases

The viral database was created by downloading protein sequences from viral genomes. Only genomes from viruses competent for infection of humans were included. Accession numbers were obtained from Entrez Genome:

(http://www.ncbi.nlm.nih.gov/sites/entrez?db=Genome), and their protein sequences retrieved using Batch Entrez (http://www.ncbi.nlm.nih.gov/sites/batchentrez) and downloaded in FASTA format. Note: Only FASTA formatted sequences can be used with command line standalone BLAST.

A FASTA formatted sequence consists of a single line called a defline, which is marked with a ">" at the beginning of the line, followed by a description of the sequence. The defline terminates with a new line, and is followed by the sequence. Sequence files exported from Entrez Batch were found to have an extra heading preceding each FASTA protein sequence, comprising the NC_xxxxx RefSeq accession of the viral genome from which the protein was derived. These lines were removed using the following command in Linux:

perl -pi -e 's/^NC_.*//s' *fastaFileName.txt*

A database of the complete human protein sequence repertoire as determined by The UniProtKB/Swiss-Prot Human Proteome Initiative (HPI) was downloaded:

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions

These sequences are in Swiss-Prot format, so they had to be converted to FASTA format. This was done using the Bioperl module, SeqIO, as follows:

```
#!/usr/bin/perl -w

   use Bio::SeqIO;


    $in = Bio::SeqIO->new(-file => "Swiss-Prot_filename" , '-format' => 'swiss');

    $out = Bio::SeqIO->new(-file => ">outputfilename.fasta" , '-format' => 'FASTA');

    while ( my $seq = $in->next_seq() ) {

        $out->write_seq($seq);

    }
```

Text files containing sequences in FASTA format cannot be used as BLAST databases directly. They must be converted to 'blastable' databases by generating BLAST indices (essentially the start position of all seed 'words' of various lengths), using the '*formatdb*' command (in more recent versions of standalone BLAST, *formatdb* has been

replaced by *makeblastdb*). There are a number of command line parameters associated with *formatdb*. These are outlined in detail in the following document available at the NCBI website:

http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html

The following command was used in Linux to format the viral and human databases:

```
formatdb -i filename.fasta -pT -oT -t "DBname"
```

-i Specifies the name of the input file containing the sequences

-p T Indicates that the input type is protein (input format T/F [true or false])

-o T Parses deflines and indexes seqIDs. Enables seqID parsing and indexing (input format T/F [true or false])

-t Adds a custom title to the database

As the human protein database for the reciprocal BLAST searches comprised sequences from Swiss-Prot, it was necessary to map each of the Entrez IDs for the differentially expressed human Treg genes and interactors, obtained from the IPA output (Section 3.2), to Swiss-Prot IDs. This was so that human query protein IDs for the first BLAST search against the viral protein database could be compared to the human protein hit IDs taken from the Swiss-Prot-derived human protein database in the reciprocal BLAST search. This was done using the database identifier mapping tool on the UniProt website:
http://www.uniprot.org/help/mapping?namespace=help&object=mapping&format=tab=batch

3.3.3 Perl and Bioperl for performing batch BLAST searches

The BLAST program, when run as a command-line utility is referred to as "*blastall*". In order to facilitate the automation of the execution of BLAST searches, a script was developed, using the Perl programming language and Bioperl modules (Appendix A). Figure 3.1 shows an overview of the workflow for the BLAST Perl script. The Bioperl project is an open source library of Perl modules for bioinformatics applications (www.bioperl.org). The Bioperl object, StandAloneBlast, which is a wrapper for the NCBI standalone BLAST package, was utilized. SearchIO which is a flexible Bioperl module for parsing pairwise alignment objects of various formats, was utilized to process the BLAST results. The parsers for BLAST output are part of SearchIO.

When using the *blastall* program/ StandAloneBlast, the BLAST parameters must first be set. There are many parameter options associated with *blastall*, details of which are available at the NCBI website:
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall.html#3

Stipulate the parameters used by the *blastall* program by populating an array. The parameters for the present study were set as follows:

@params_virus = (-p => 'blastp', -d => 'viral_DB.fasta', -o => 'report.bls', -e => '10')

-p Specifies the type of search. In this case, blastp, protein querying a protein database.

-d Specifies the target database(s)

-o Specifies output file

-e Specifies Expectation value cutoff (default is ‘10’)

-b Number of database sequence alignments shown. Set to ‘1’ so only top hit is returned

-v Number of one line description of database sequences shown

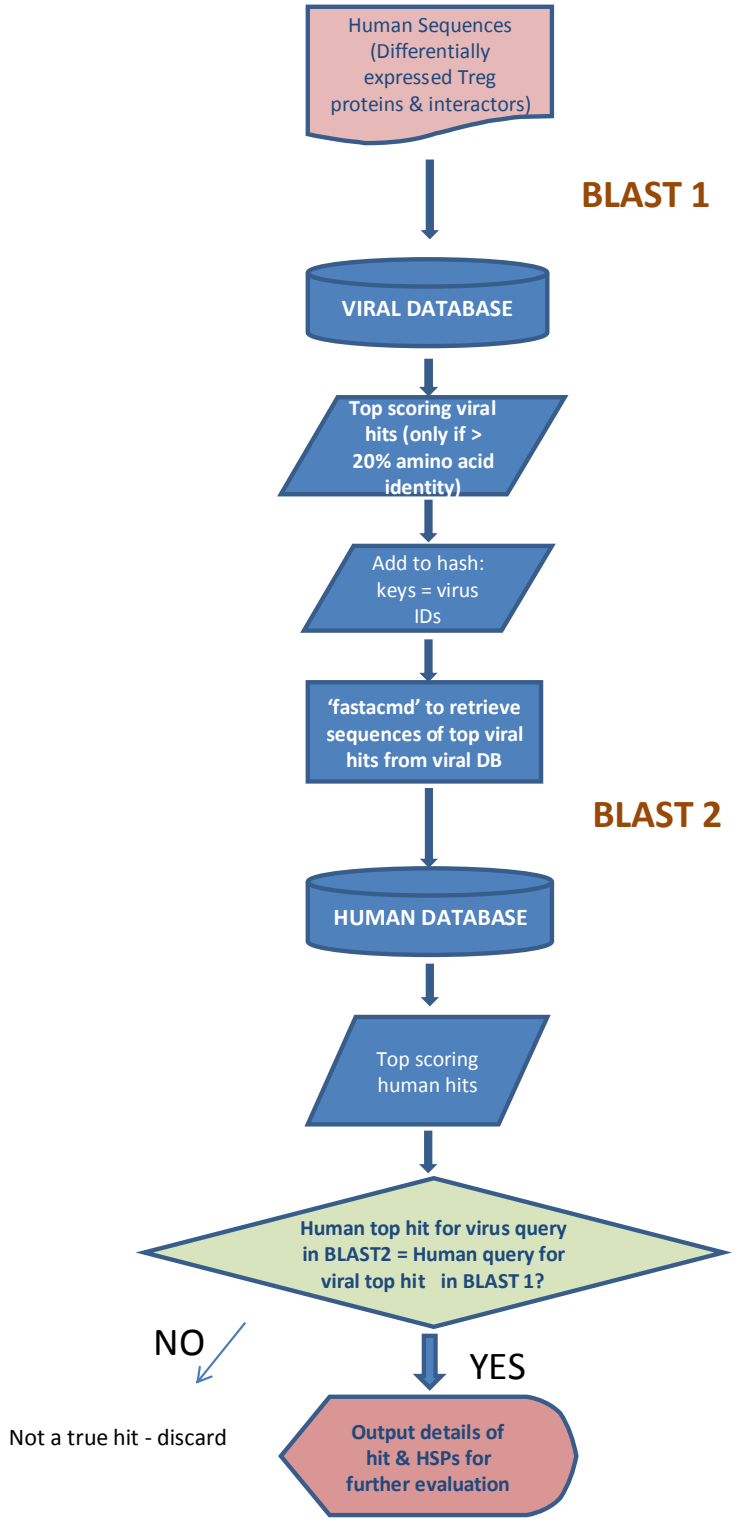


Figure 3.1 Workflow for batch reciprocal BLAST

The query sequences for the first BLAST search were contained in a single file of all the sequences of differentially expressed Treg proteins and their interacting proteins. A Bio::Seq object was instantiated for each query sequence in the file. A StandAloneBlast object (a.k.a "factory") was created, and *blastall* executed by calling the *blastall* method with the blast parameter array. The parsed BLAST output is returned as a BLAST report which is a Bio::SearchIO object. SearchIO is used to extract data from the BLAST report. Each BLAST result is composed of a set of hits for the query sequence. Hits are sequences in the searched database which align with the query sequence while meeting defined search parameters e.g maximum E-value. Each hit has one or more high-scoring segment pairs (HSPs), which are significant alignments of the query and hit sequence, where significance is defined by E value thresholds. The significance of a hit is ascribed to the E value of the highest scoring HSP. The search results are therefore made up of three main components: The top level results, the hits, and the HSPs. Each of these outputs have methods associated with them which can be used to obtain information about the BLAST results (Tables 3.1-3.3). The required information on the result, hits or HSPs can be specified and output to the screen or saved as a file.

In order to retrieve the full sequence of all hits for subsequent reciprocal Blasting, the '*fastacmd*' command must be used. This is another database-related tool from the standalone BLAST package. It allows use of formatted BLAST databases for non-sequence alignment purposes, such as dumping of FASTA sequences, extracting information for specific entries, and retrieving specific sequences or subsequences. Retrieving specific entries requires the database to be formatted with "-o T" (see section 3.4.3 above). *fastacmd* was used to retrieve the sequences of the top viral hits in the first BLAST search, which were then used as input for the second, reciprocal BLAST search.

Table 3.1 Methods associated with 'Result' objects (from: http://bioperl.org/wiki/HOWTO:SearchIO)

Object	Method	Example	Description
Result	algorithm	BLASTX	algorithm string
Result	algorithm_version	2.2.4 [Aug-26-2002]	algorithm version
Result	query_name	20521485|dbj|AP004641.2	query name
Result	query_accession	AP004641.2	query accession
Result	query_length	3059	query length
Result	query_description	Oryza sativa ... 977CE9AF checksum.	query description
Result	database_name	test.fa	database name
Result	database_letters	1291	number of residues in database
Result	database_entries	5	number of database entries
Result	available_statistics	effectivespaceused ... dbletters	statistics used
Result	available_parameters	gapext matrix allowgaps gapopen	parameters used
Result	num_hits	1	number of hits
Result	hits		List of all Bio::Search::Hit::GenericHit object(s) for this Result
Result	rewind		Reset the internal iterator that dictates where next_hit() is pointing, useful for re-iterating through the list of hits

Table 3.2 Methods associated with 'Hit' objects (from: http://bioperl.org/wiki/HOWTO:SearchIO)

Object	Method	Example	Description
Hit	name	443893|124775	hit name
Hit	length	331	Length of the Hit sequence
Hit	accession	443893	accession (usually when this is a genbank formatted id this will be an accession number- the part after the *gb* or *emb*)
Hit	description	LaForas	hit description

		sequence	
Hit	algorithm	BLASTX	algorithm
Hit	raw_score	92	hit raw score
Hit	significance	2e-022	hit significance
Hit	bits	92.0	hit bits
Hit	hsps		List of all Bio::Search::HSP::GenericHSP object(s) for this Hit
Hit	num_hsps	1	number of HSPs in hit
Hit	locus	124775	locus name
Hit	accession_number	443893	accession number
Hit	rewind		Resets the internal counter for next_hsp() so that the iterator will begin at the beginning of the list

Table 3.3 Methods associated with 'HSP' objects (from: http://bioperl.org/wiki/HOWTO:SearchIO)

Object	Method	Example	Description
HSP	algorithm	BLASTX	algorithm
HSP	evalue	2e-022	e-value
HSP	expect	2e-022	alias for evalue()
HSP	frac_identical	0.884615384615385	Fraction identical
HSP	frac_conserved	0.923076923076923	fraction conserved (conservative and identical replacements aka "fraction similar") (only valid for Protein alignments will be same as frac_identical)
HSP	gaps	2	number of gaps
HSP	query_string	DMGRCSSG ..	query string from alignment
HSP	hit_string	DIVQNSS ...	hit string from alignment

HSP	homology_string	D+ + SSGCN ...	string from alignment
HSP	length('total')	52	length of HSP (including gaps)
HSP	length('hit')	50	length of hit participating in alignment minus gaps
HSP	length('query')	156	length of query participating in alignment minus gaps
HSP	hsp_length	52	Length of the HSP (including gaps) alias for length('total')
HSP	frame	0	$hsp->query->frame,$hsp->hit->frame
HSP	num_conserved	48	number of conserved (conservative replacements, aka "similar") residues
HSP	num_identical	46	number of identical residues
HSP	rank	1	rank of HSP
HSP	seq_inds('query','identical')	(966,971,972,973,974,975 ...)	identical positions as array
HSP	seq_inds('query','conserved-not-identical')	(967,969)	conserved, but not identical positions as array
HSP	seq_inds('query','conserved')	(966,967,969,971,973,974,975, ...)	conserved or identical positions as array
HSP	seq_inds('hit','identical')	(197,202,203,204,205, ...)	identical positions as array
HSP	seq_inds('hit','conserved-not-identical')	(198,200)	conserved not identical positions as array
HSP	seq_inds('hit','conserved',1)	(197,202-246)	conserved or identical positions

			as array, with runs of consecutive numbers compressed
HSP	score	227	score
HSP	bits	92.0	score in bits
HSP	range('query')	(2896,3051)	start and end as array
HSP	range('hit')	(197,246)	start and end as array
HSP	percent_identity	88.4615384615385	% identical
HSP	strand('hit')	1	strand of the hit
HSP	strand('query')	1	strand of the query
HSP	start('query')	2896	start position from alignment
HSP	end('query')	3051	end position from alignment
HSP	start('hit')	197	start position from alignment
HSP	end('hit')	246	end position from alignment
HSP	matches('hit')	(46,48)	number of identical and conserved as array
HSP	matches('query')	(46,48)	number of identical and conserved as array
HSP	get_aln	*sequence alignment*	Bio::SimpleAlign object

3.3.4 Evaluation of aligned sequences

Each BLAST result is associated with an E-value (Expect value), which represents the number of times a hit of the same score or better would be expected to occur purely by chance. This value is dependent on the lengths of the query sequence and search space (database), as well as the identity between the query and target sequences. The lower the E-value, the greater the similarity between the input sequence and the match. The biological significance of the viral and human sequence alignments was also examined. This was undertaken by manually examining the sequence alignments. In particular, by examining the alignment of cysteine residues in the sequences where they were present. Cysteines form disulphide bridges and so are important in maintaining the tertiary structure of proteins. Mismatches in cysteine content between the human protein and its viral homologue may therefore indicate differences in protein folding and function. In addition, the sequences were examined for the presence of shared motifs/domains. If the viral and human protein shared common motifs or domains within their sequences, this could indicate shared biological function or membership of similar protein families. The presence of motifs and signature domains was examined using the InterproScan database, at the EBI:

http://www.ebi.ac.uk/Tools/pfa/iprscan/

3.4 Identification of putative functional modules

A Treg protein interaction network with confidence scores for each interaction, and putative active subnetworks were generated as follows:

(i) 301 differentially expressed Treg proteins comprising the consensus Treg signature (see section 3.1) were used to query public PPI databases and generate a custom protein interactome.

(ii) Confidence scores for each of these binary interactions were obtained.

(iii) The Treg gene expression data was overlaid onto the PPI interactome/network.

(iv) Functional modules or active subnetworks were identified.

3.4.1 Protein-protein interaction data sources

There are many publically available molecular interaction databases; however there is often little overlap in their data content (Turinsky *et al.*, 2011). Therefore, in order to generate a comprehensive PPI network, PPI data was obtained from a number of different sources. The PSI (Proteomics Standards Initiative) common query interface (PSICQUIC) is an online meta-search resource which allows multiple interaction data sources to be queried simultaneously (Aranda *et al.*, 2011):

http://www.ebi.ac.uk/Tools/webservices/psicquic/view

A single query is sufficient to retrieve the relevant interaction data from all of the interaction data sources. PSICQUIC utilizes the Molecular Interaction Query Language (MIQL), which allows searches for specific organisms, interaction detection methods, interaction types, or publication identifiers. Search results may be clustered if required. This will remove redundant interactions from the different source databases, i.e. if the same binary interaction between two proteins is found in more than one database, the results will be merged into a single interaction.

PSICQUIC sources its data from a number of different databases. Only high-quality, manually curated data were to be included in the PPI network. Therefore, rather than clustering the results, only data from databases which were selected based upon their coverage, content and data curation practices were exported from PSICQUIC. Data was downloaded from results returned from the IntAct, MINT, DIP, MatrixDB and APID data repositories. Results are exported from PSICQUIC in the PSI-MI format, a community standard for the representation of molecular interaction data, introduced by the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI). Two PSI-MI compliant formats are supported: PSI-MI XML (PSI molecular interaction XML) or PSI-MI TAB (PSI molecular interaction tabular) formats (only MI-TAB if more than 200 binary interactions). The column contents of the MI-TAB format are as follows:

1. **Unique identifier for interactor A**, represented as databaseName:ac, where databaseName is the name of the corresponding database as defined in the PSI-MI controlled vocabulary, and ac is the unique primary identifier of the molecule in the database. Identifiers from multiple databases can be separated by "|". It is recommended that proteins be identified by stable identifiers such as their UniProtKB or RefSeq accession number.
2. **Unique identifier for interactor B**.
3. **Alternative identifier for interactor A**, for example the official gene symbol as defined by a recognised nomenclature committee. Representation as databaseName:identifier. Multiple identifiers separated by "|".
4. **Alternative identifier for interactor B**.
5. **Aliases for A**, separated by "|". Representation as databaseName:identifier. Multiple identifiers separated by "|".
6. **Aliases for B**.
7. **Interaction detection methods**, taken from the corresponding PSI-MI controlled Vocabulary, and represented as databaseName:identifier(methodName), separated by "|".
8. **First author** surname(s) of the publication(s) in which this interaction has been shown, optionally followed by additional indicators, e.g. "Doe-2005-a". Separated by "|".
9. **Identifier of the publication** in which this interaction has been shown. Database name taken from the PSI-MI controlled vocabulary, represented as databaseName:identifier. Multiple identifiers separated by "|".
10. **NCBI Taxonomy identifier for interactor A**. Database name for NCBI taxid taken from the PSI-MI controlled vocabulary, represented as databaseName:identifier. Multiple identifiers separated by "|". Note: In this column, the databaseName:identifier(speciesName) notation is only there for consistency. Currently no taxonomy identifiers other than NCBI taxid are

anticipated, apart from the use of -1 to indicate "in vitro" and -2 to indicate "chemical synthesis".

11. **NCBI Taxonomy identifier for interactor B**.
12. **Interaction types**, taken from the corresponding PSI-MI controlled vocabulary, and represented as dataBaseName:identifier(interactionType), separated by "|".
13. **Source databases** and identifiers, taken from the corresponding PSI-MI controlled vocabulary and represented as databaseName:identifier(sourceName). Multiple source databases can be separated by "|".
14. **Interaction identifier(s)** in the corresponding source database, represented by databaseName:identifier
15. **Confidence score**. Denoted as scoreType:value. There are many different types of confidence score, but so far no controlled vocabulary. Thus the only current recommendation is to use score types consistently within one source. Multiple scores separated by "|".

All columns are mandatory, although they will not all necessarily contain data values.

The UniProt IDs for all of the differentially regulated Treg proteins comprising the consensus Treg signature were pasted into the search area. The results from each of the APID, DIP, IntAct, MatrixDB and MINT databases were exported as PSI MI-TAB formatted files.

3.4.2 Obtaining confidence scores for PPIs

The PSI confidence scoring system (PSISCORE) was developed to provide a means for assessing the quality and reliability of molecular interaction data. There are numerous methods for determining PPIs, each with their own pros and cons. For example there are many different experimental techniques producing results of varying reliability, from high-throughput yeast 2-hybrid studies through to small-scale single protein studies, as well as methods utilizing computational predictions. This makes it difficult

to objectively assess the relative strengths of evidence supporting each interaction. To date, there is no community-consensus scoring scheme for molecular interactions. PSISCORE is a decentralized system, which utilizes individual servers that apply different confidence scoring methods to interaction data. The user can select which scoring method(s) to apply (Figure 3.2).

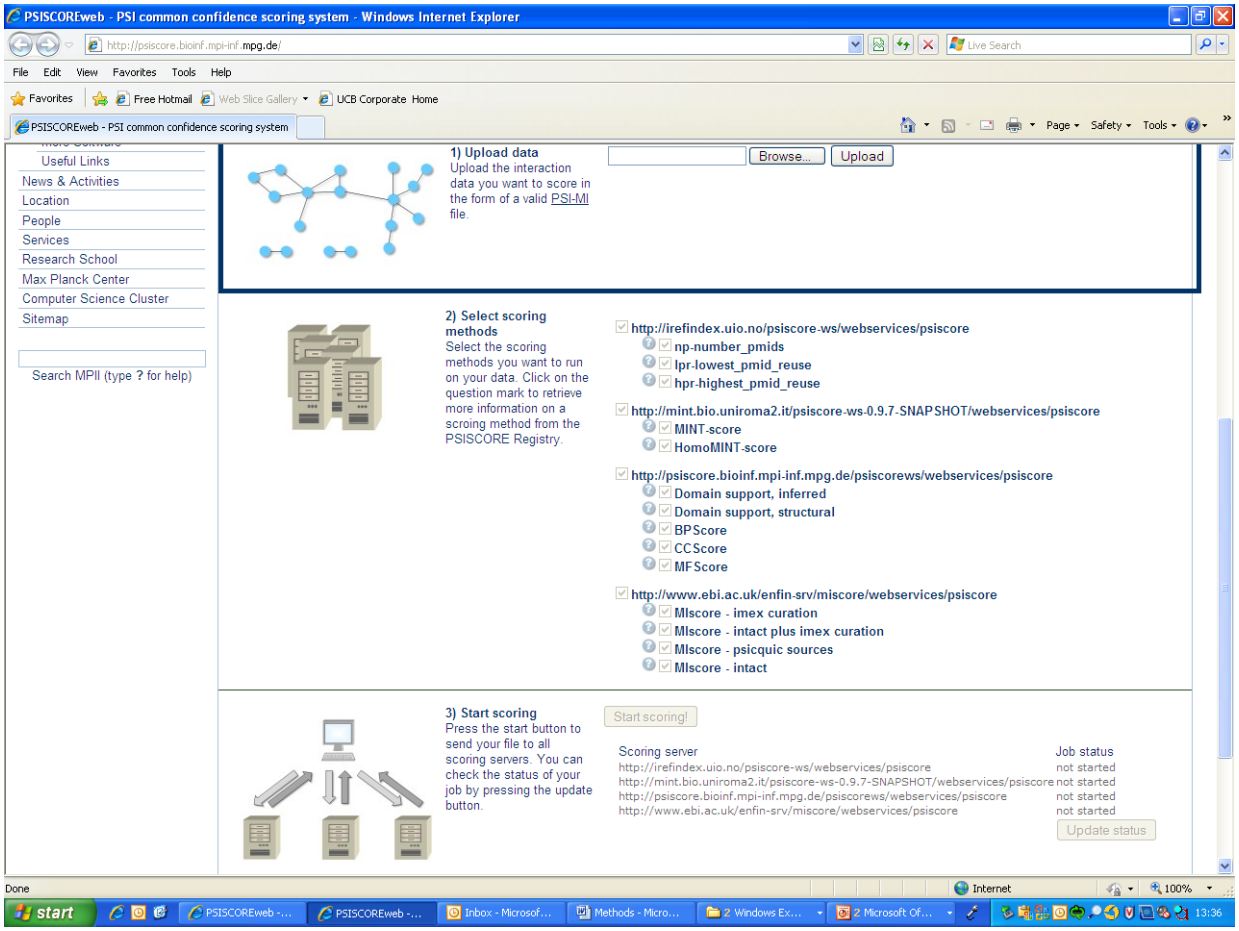


Figure 3.2 Screenshot showing the PSISCORE web interface and different scoring options

The different scoring options are as follows (descriptions taken from the PSCISCORE project website: http://code.google.com/p/psiscore/wiki/Scoring_methods_overview and the PSISCORE web viewer: http://psiscore.bioinf.mpi-inf.mpg.de/):

iRefIndex scores:

iRefIndex calculates three scores based on the publications that provide evidence for an interaction.

lpr (lowest pmid re-use) defines the lowest number of distinct interactions that a publication supporting the interaction is used to support. Low values (e.g. one) indicate that at least one of the publications that support an interaction has only reported few (for one this means no other) interactions, which is likely the case for low-throughput experiments. Range: 0-inf

hpr (highest pmid re-use) is the highest number of interactions that a publication supporting the interaction is used to support. High values indicate that a publication describes many other interactions as well, which is likely the case for high-throughput methods. Range: 0-inf

np (number pmids) is the total number of unique publications used to support the interaction. Higher values indicate that an interaction has been reported in multiple publications.

Range: 0-inf

MINT score:

The MINT score takes into account experimental evidence associated with the interaction detection method.

Range: 0-1

PSISCORE:

Domain Support – inferred:

Domain support indicates protein-protein interactions that can be traced to the underlying protein domain-domain interactions (DDIs). 'Domain support, inferred' contains DDIs from several computational predictions.

Range 0-1

Domain Support – structural:

Domain support indicates protein-protein interactions that can be traced to the underlying protein domain-domain interactions. 'Domain support, structural' contains DDIs from datasets that have been inferred from structural information.

Range 0-1

BP Score:

The BPscore is a measure of the functional similarity between two proteins or protein families with respect to their biological process annotation of the Gene Ontology.

Range 0-1

MF Score:

The MFscore is a measure of the functional similarity between two proteins or protein families with respect to their molecular function annotation of the Gene Ontology.

Range 0-1

MI score:

This score is designed to calculate annotation evidence based on common and minimun curated information reporting a molecular interaction experiment.

The result is a score between 0 and 1 per interaction. It will take into account several variables:

- Number of publications
- Experimental detection methods found for the interaction
- Interaction types found for the interaction

Each of these variables will be represented by a score between 0 and 1. The importance of each variable in the equation will be adjusted using a weight factor. The publication score will take into account the different publications supporting the interaction. The method score will be calculated using the MI ontology (experimental interaction detection terms) giving preference in the equation to terms with higher assigned values. The method score will also take into account the diversity of methods reported for the interaction.

Range: 0-1

The MIscore scoring method was utilized in this study. This takes into account data derived from all PSICQUIC data sources for a particular interaction when calculating the confidence score.

The input to PSISCORE must be a HUPO-PSI-defined file format (i.e. PSI-MI TAB or PSI-MI XML). This file is then sent to the appropriate scoring servers, depending on the scoring method(s) selected. All the calculated scores are then added to the initial input file (added to column 15 in the case of a MI TAB file).

The PSI-MI TAB files exported from PSICQUIC are each individually submitted to the PSISCORE web client. In instances where the file was large, it was first split into smaller files of ≤ 2000kb using the Linux 'split' command (PSISCORE states it will currently not handle files larger than 5120kb, although the size limit was found to be actually less than half this):

```
split -b 2000k big_file.txt new_file
```

This will split the file 'big_file.txt' into smaller files, each of size 2000kb with the names new_fileaa, new_fileab, new_fileac…..etc. These were later joined back together after scoring, using the Linux 'cat' function:

```
cat scored_new_fileaa.txt scored_new_fileab.txt
>joined_scored_file.txt
```

The scored interactions were appended as a new column (column 15) to the input MI-TAB file. It should be noted that many interactions were returned with no score associated with them, or a score of zero was returned. The PSISCORE developers and administrators of the MIscore server have been informed, and the reasons for this have, as yet, not been determined.

Although all data was returned in the PSI-MI TAB format, the formats for the annotation of the molecules in columns 1 and 2 varied, depending on the source database from which the data was obtained. Columns 1 and 2 were formatted as follows:

APID and MINT:

uniprotkb:P78540

IntAct:

uniprotkb:Q13451|intact:EBI-306914

DIP:

dip:DIP-38276N|uniprotkb:P54259

MatrixDB:

uniprotkb:P16109|matrixdb:P16109

To enable overlaying and integration of other data-types in Cytoscape later (Section 3.4.5), each of these columns (which will later be represented as nodes in the Cytoscape network) must be in the same format, i.e. possess a common identifier. Columns 1 and 2 for each of these exported MI TAB files possess a UniProt ID (along with other identifiers in the DIP, IntAct and MatrixDB results), so this was selected as the common identifier.

Column 12, contains a description of the interaction type in the following format:

psi-mi:"MI:0915"(physical association)

Column 15, added by PSISCORE, contains the confidence score in the following format:

MIscore:0.40116468(psicquic sources including APID,ChEMBL,BioGrid,IntAct,DIP,InnateDB,MPIDB,iRefIndex,MatrixDB,MINT,Interoporc,Reactome,Reactome-FIs,STRING,BIND,DrugBank)

Scripts were therefore written in Perl to parse the UniProt IDs, as well as the interaction types and confidence scores for clearer representation in the Cytoscape network later (Appendix B).

3.4.3 Generation of a scored PPI network

The scored interaction data were visualized using Cytoscape, an open source software platform for visualizing complex networks and integrating these with any type of attribute data (Shannon *et al.*, 2003). It is available as a platform-independent open-source Java application, and must be downloaded onto a local computer from the Cytoscape website (www.cytoscape.org).

Cytoscape is able to import data in a number of different formats, including the PSI-MI formats. The PSI-MI TAB-formatted scored interaction data were imported into Cytoscape as a 'Network from a table' (Figure 3.3). Columns 1 and 2 were selected as the source and target interactions respectively, and column 12 was selected as the interaction type. Column 15 was selected as an edge attribute.

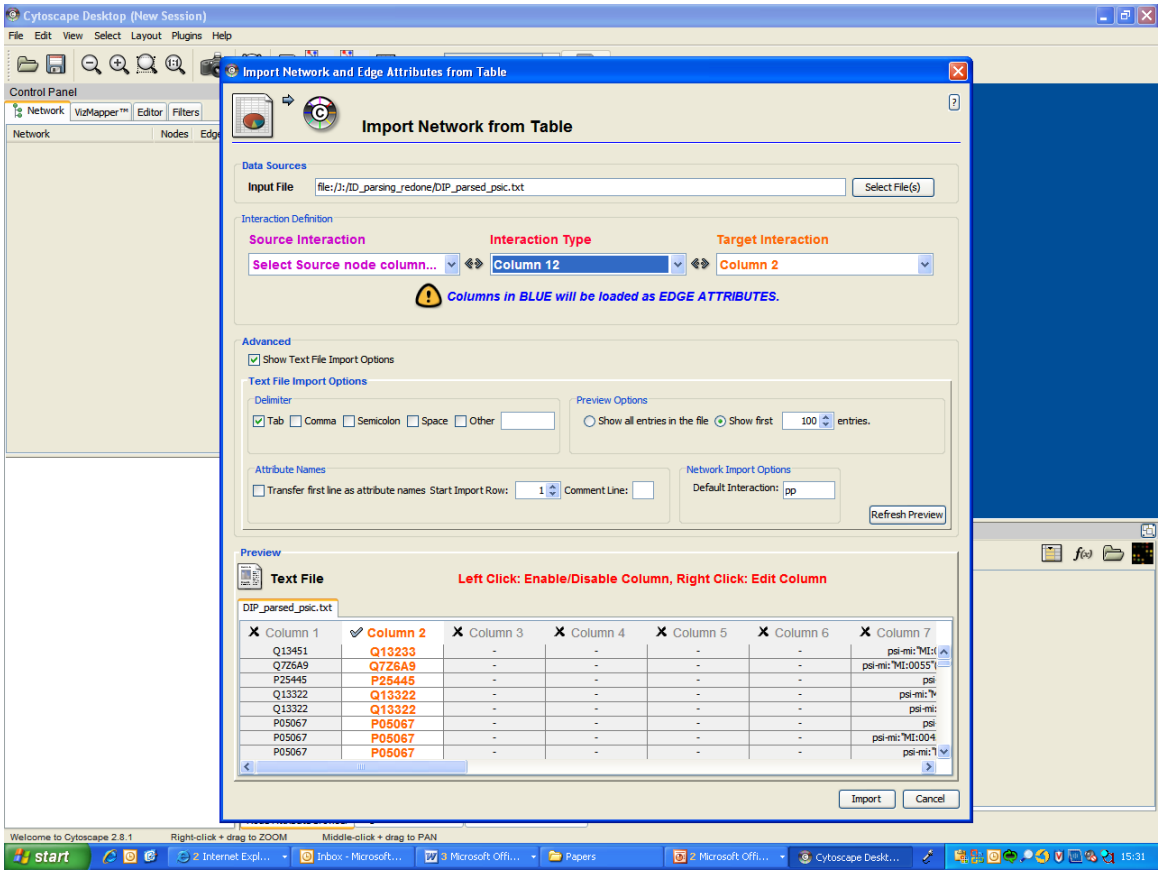


Figure 3.3 Screenshot taken from Cytoscape, showing import of scored interaction data in PSI-MI TAB format

Once the network has been imported, proteins are represented as the nodes, and the interaction relationships between them are represented by the edges, i.e. lines connecting them (a user manual for Cytoscape is available at www.cytoscape.org). Clicking on an edge and using the edge attribute viewer in the data panel will display the confidence score value. At this stage, the PPI network was represented as five separate networks, one for each of the imported MI TAB files. There was overlap between these networks for many of the interactions, i.e. the same binary interaction may be present in more than one database. In order to obtain a non-redundant PPI network, the core Cytoscape plugin, 'advanced network merge' was applied to the networks, using the 'union' operation (Figure 3.4). This resulted in the creation of a single large network. Incorporating interactions from all databases, but with duplicate interactions removed. Duplicate edge removal by Cytoscape is based upon the

interaction attributes (column 12 in the MI TAB file). If two interactions are comprised of two identical nodes joined by an identical interaction type it will be removed. This, of course is very much dependent upon the interaction type description attributed to the data in the original data source, i.e. unified interaction vocabulary between data sources is necessary. If one data source describes an interaction as 'physical interaction' and another describes the same interaction as 'association', these will be treated as two distinct binary interactions and both will be retained. There is therefore the possibility that some duplicate interactions may be retained after network merge.

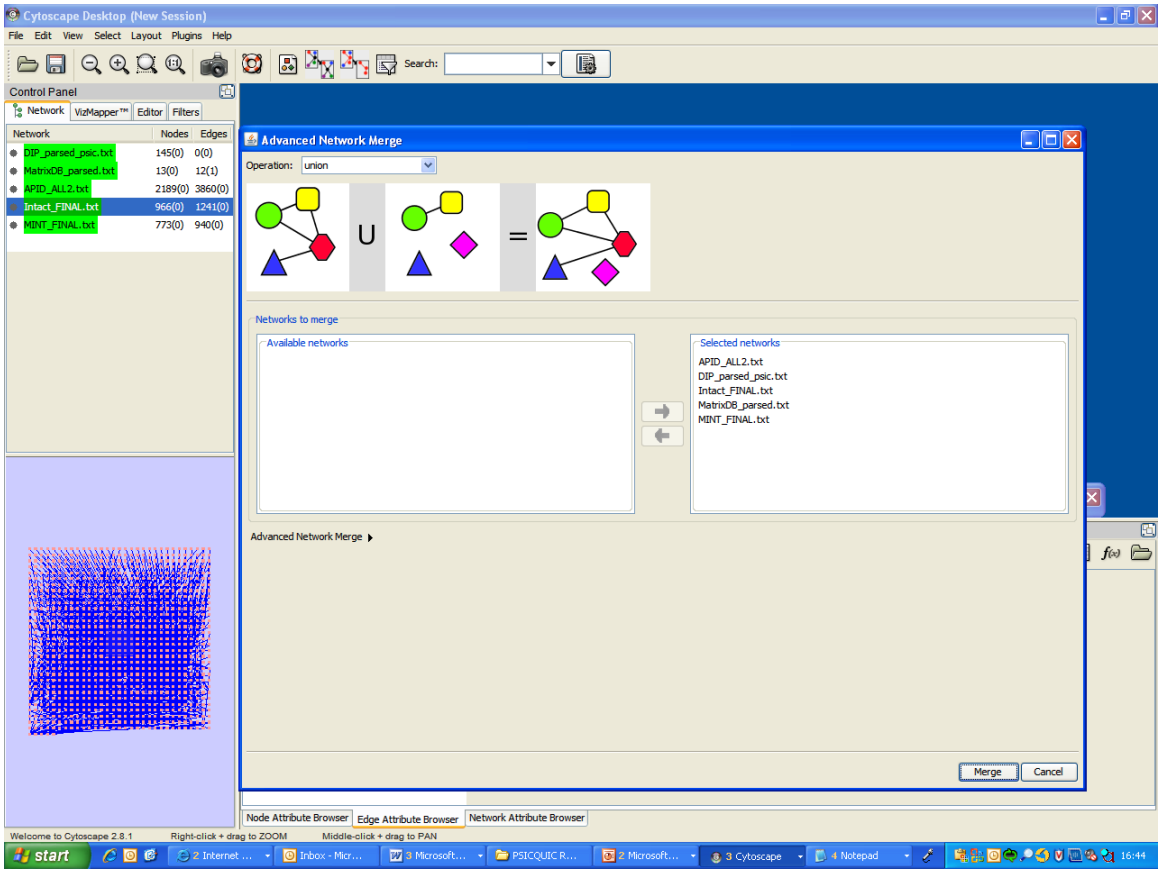


Figure 3.4 Screenshot taken from Cytoscape, showing advanced network merge

3.4.4 Identification of putative functional modules

Functional modules or subnetworks can be identified as highly inter-connected network regions enriched for genes that show significant changes in gene expression in Tregs. Putative functional modules were identified using the Cytoscape plugin, jActiveModules, which uses algorithms to search for connections between gene expression and network topology (Ideker *et al.*, 2002). The plugin can be installed via the plugin manager in Cytoscape.

In order to import gene expression data into Cytoscape, the gene or protein identifier in the data file must exactly match the corresponding, previously loaded Cytoscape node identifier. The nodes identifiers in the PPI network generated in section 3.4.4 were UniProt IDs. It was therefore necessary to map the Entrez gene IDs for all the consensus, differentially expressed Treg genes to the corresponding UniProt IDs. The jActiveModules plugin will only recognize p-values associated with gene expression data. The mean p-value across the datasets from which the consensus signature was calculated for each gene. Cytoscape will only import expression data if it is in a specified format. The data must be organized as a matrix, with each row representing the expression data for one gene. The first row must provide column labels. Column 1 must hold the gene identifier, column 2 must hold arbitrary text such as a description, column 3 and subsequent columns contain expression values. If the data matrix is created in a spreadsheet program such as Excel, the columns must be merged into a single cell, and the file saved as a text file, having the extension either .mRNA or .pvals. The Treg gene expression data file format looked something like this:

```
Uniprot_ID Gene_symbol Av_FC Av_pval
Q5JTZ9  AARS2 -1.715 0.01072425
P08183  ABCB1 -2.55 0.023475
O14639  ABLIM1 -0.658 0.01436
```

Gene expression data was uploaded using the Cytoscape ‘Import expression/attribute matrix’ option.

Once the gene expression data had been imported, jActiveModules plugin was utilized to commence the search for putative functional modules.

Parameters in the General Parameters panel of JActiveModules are:

“

(a) Number of modules – indicates the number of putative modules that will be reported

(b) Adjust score for size – corrects for the fact that a larger module is more likely to contain nodes with significant P-values by random chance

(c) Regional scoring – Instead of scoring only those nodes within the module, the neighbouring nodes of the module are also included.

Parameters in the Strategy Panel:

(a) Search – local (greedy) searches are initiated from single nodes in the network

(b) Search depth – at each step in the greedy search, this parameter determines how close a node must be to the current active module to be considered for inclusion

(c) Max depth – determines how close a node must be to the initial seed node to be considered for inclusion

(d) Search from selected nodes – by default, a separate search is initiated for each node in the network. Using this option, searches are initiated only from nodes selected by the user

(e) Anneal – all active modules are discovered simultaneously using the method of simulated annealing (Ideker *et al.*, 2002).

”

Descriptions taken from (Cline *et al.*, 2007)

.

The ‘search’ strategy was selected in the General Parameters panel. The nodes representing the significant viral hits (identified using the strategies described above in sections 3.1 - 3.3) were used as seed nodes from which to initiate the search, as

information on potential biological functions and significance of those proteins/genes were of most interest. The other parameters were initially kept at the default values (Adjust score for size, Regional scoring, Search depth = 1, Max depth = 2), with the search depths being adjusted for subsequent runs. JActiveModules finds the highest-scoring networks (i.e. the highest-scoring connected subgraphs). Network scores are aggregate Z-scores based on p-values (The Z-score calculation is described in detail by Ideker *et al.*, 2002). The results panel for each run contained a table, with each row representing a putative module and an associated Z-score. Higher Z-scores indicate biologically active networks. Z-scores greater than 3 were considered significant.

BiNGO (Biological Networks Gene Ontology tool) is a tool, available as a Cytoscape plugin, to determine which Gene Ontology (GO) categories are statistically overrepresented in a set of genes or a subgraph of a biological network (Maere *et al.*, 2005). BiNGO was used to assess if the nodes in the functional modules were enriched for biological processes recorded in the GO database (Ashburner *et al.*, 2000).

4 Results

In the following sections, there will be reference to four genes that are under investigation as potential new targets by UCB. These data and the identities of these genes hereby anonymized as 'Tolerogenic Genes Of Interest-1, 2, 3 and 4' (TGOI-1, TGOI-2, TGOI-3, TGOI-4).

The methodology utilized and the numbers of genes/proteins returned at each stage of the process are summarized in figure 4.1.

4.1 Generation of a consensus Treg gene signature

NextBio consists of a library of gene expression studies, each comprising one or more biosets. A bioset is a set of differentially expressed gene data derived from a single experiment. A study is a set of biosets that correspond to a single published research paper.

Nine studies of potential interest were found in NextBio, consisting of a total of twelve biosets. These were studies comparing gene expression in human Treg cells v naïve or conventional T cells, as defined by the study authors. Six studies were sourced from GEO, two from Array Express, and one from MACE (transcriptome repository, IHES, France). Of these twelve biosets, six fulfilled the selection criteria for inclusion in the analysis, i.e. showed increased expression in Treg cells of the pre-determined combination of key Treg-associated genes.

There were a total of 303 genes that were differentially expressed in Treg cells v non-Treg cells in at least four out of the six studies. 117 were upregulated in Treg cells

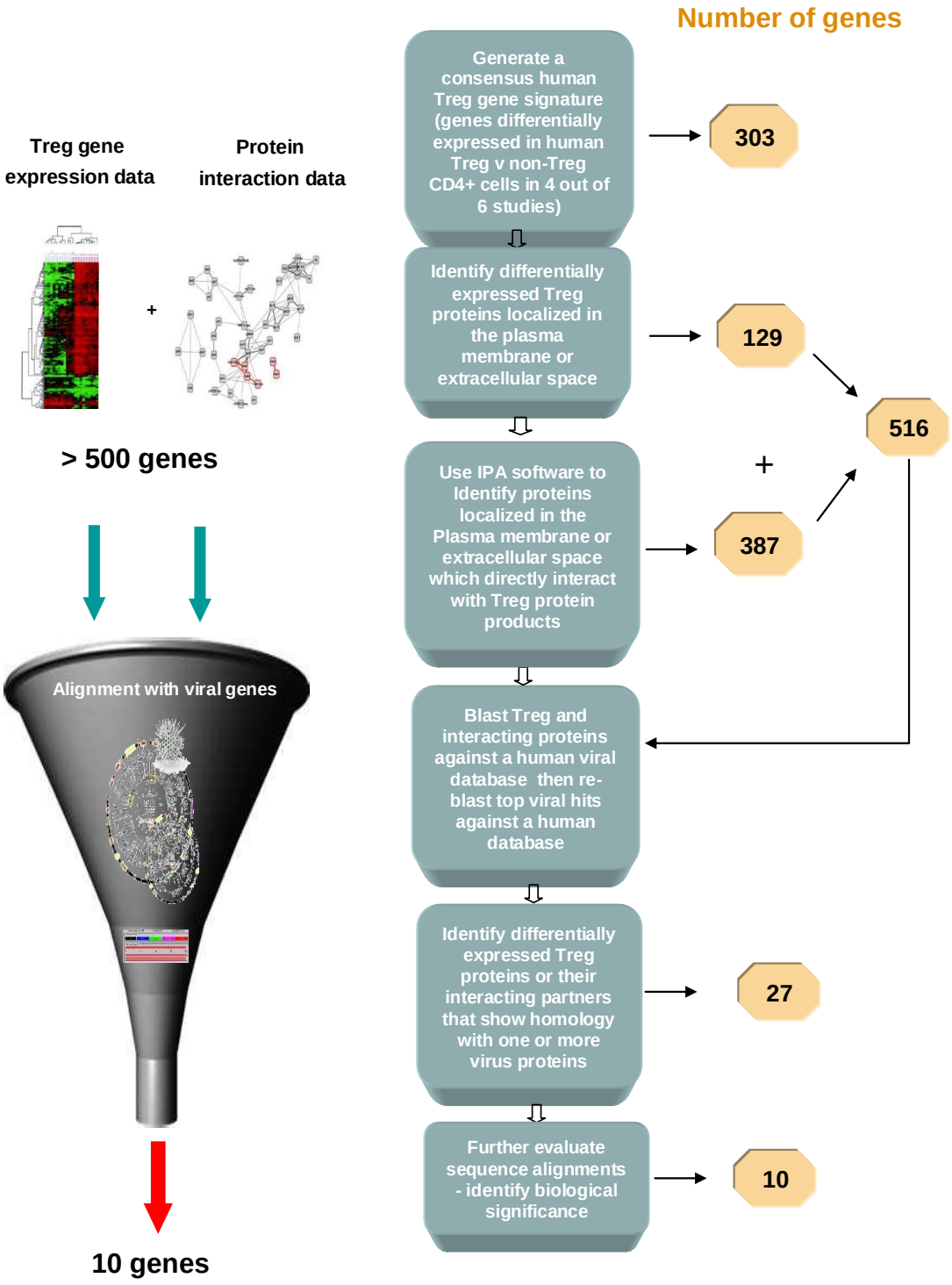


Figure 4.1 Summary of the workflow and processes used to identify putative tolerogenic genes

compared to non-Treg cells, 186 were downregulated, i.e. showed higher expression in non-Treg cells.

4.2 Expansion of the Treg gene signature to include interacting proteins

The 303 differentially expressed Treg genes were uploaded into IPA. Four of these gene identifiers were not mapped by IPA to a molecule in the IPA database. Two of these genes were manually re-annotated and were subsequently successfully mapped by IPA. However, two genes remained unmapped. These were FLJ23834, a hypothetical protein, and RP6-213H19.1, also known as serine threonine protein kinase MST4. MST4 and its associated interacting proteins were present in the IPA database, but it was unmapped for unknown reasons. MST4 was therefore analysed separately within IPA, and being a nuclear protein was subsequently eliminated from the analysis. 301 mapped genes were analysed within IPA. Of these 301 gene products, 65 were localized in the nucleus, 107 in the cytosol, 67 in the plasma membrane (PM), 16 in the extracellular space and 46 were of unknown localization (Figure 4.2). Proteins localized in the PM and extracellular space were of interest, but proteins in unknown locations were also retained to ensure that nothing of potential interest but of unknown localization was discarded. This gave a total of 129 Treg proteins of interest. 387 proteins localized in the plasma membrane, extracellular space or in unknown locations were found to directly interact with the differentially expressed Treg gene products, according to the filtering criteria described in the methods (Figure 4.3). The 129 Treg proteins plus 387 interacting proteins gave a total of 516 proteins to take forward to the next stage of the analysis.

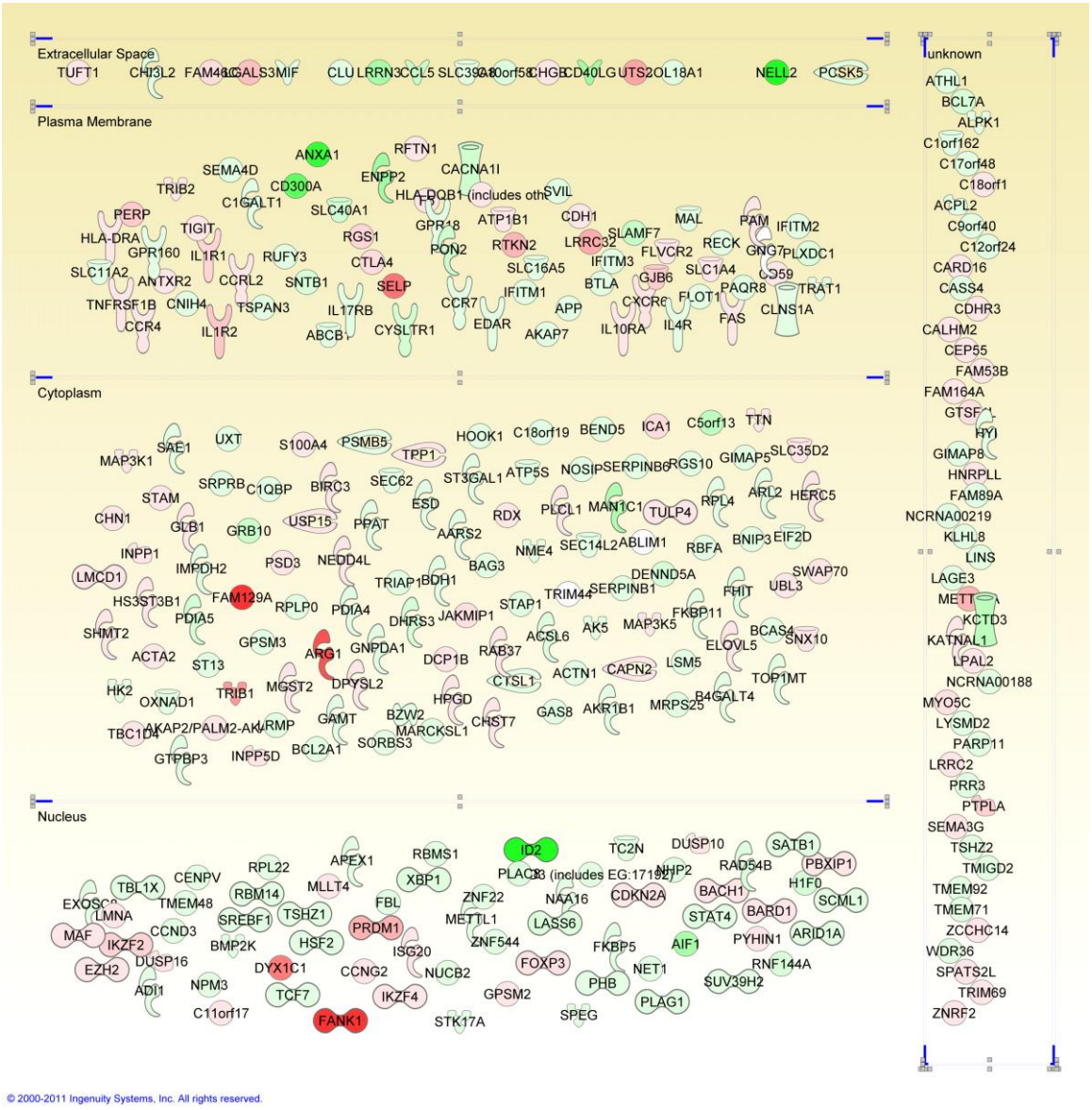


Figure 4.2 The 301 differentially expressed Treg proteins – subcellular localizations

Red = increased gene expression in Treg cells, Green = decreased gene expression in Treg cells. Intensity of colour is proportional to level of gene expression – darker colour = bigger change in expression

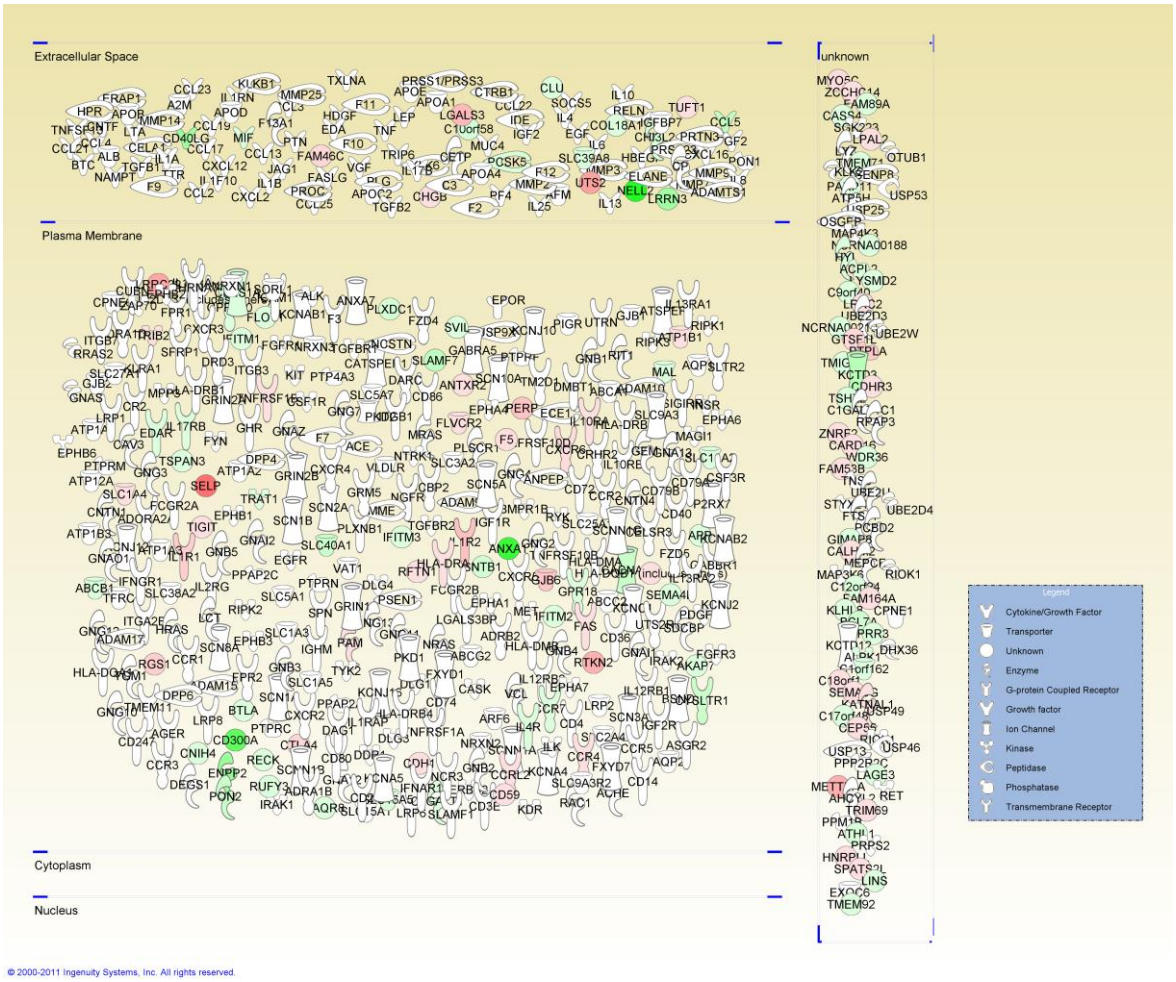


Figure 4.3 The 516 differentially regulated Treg proteins + direct interactors.

Red = increased gene expression in Treg cells, Green = decreased gene expression in Treg cells. Intensity of colour is proportional to level of gene expression – darker colour = bigger change in expression

4.3 Search for viral homologues

The presence of differentially expressed Treg genes in viral genomes was used as a filter for identifying putative tolerogenic genes (Figure 4.1). The first BLAST search used the 516 differentially expressed human Treg gene products and their interactors as input sequences for alignment against a database of viral proteins. 503 of the human proteins aligned with one or more viral proteins. When the top scoring viral hits for each of the human proteins were re-BLASTED back against the database of human protein sequences, only 37 human hits were returned which were identical to the original query sequence in the initial BLAST. This number was further reduced to 27 hits, as 10 of these proteins, after advice from UCB experts in the field, were deemed unsuitable as targets and were excluded from the list (Table 4.4). These 27 human proteins with homologous viral sequences could be grouped into four categories: Upregulated expression in Treg cells (3 fell into this category) [Figure 4.4], direct interactor with an upregulated Treg protein (12) [Figure 4.5], downregulated expression in Treg cells (12) [Figure 4.6], and direct interactor with a downregulated Treg protein (4) [Figure 4.7]. Some proteins fell into more than one category, i.e. they interacted with both downregulated and upregulated Treg proteins: C3, ITGB1 and TGOI-3.

4.4 Evaluation of aligned sequences

The E-value associated with each BLAST result is a measure of statistical significance, representing the number of times the match would be expected to occur purely by chance in a search of a database of a particular size. Statistical significance, however, does not always equate to biological significance; in this case preserved function. Therefore, further analysis of the alignments to examine the biological significance of the viral and human sequence alignments was also undertaken. The alignments of cysteine residues in the sequences were inspected. Cysteines form disulphide bridges and so are important in maintaining the tertiary structure of proteins. Mismatches in cysteine content between the human protein and its viral homologue may therefore indicate differences in protein folding and function. The aligned human and viral sequences were also examined for the presence of shared motifs and signature domains.

If the viral and human protein shared common motifs or domains within their sequences, this could indicate shared biological function or membership of the same protein family.

E values, conservation of cysteine residues and shared protein family domains were considered collectively to evaluate whether a human/virus protein alignment was a significant hit. Table 4.5 shows the results of these assessments - favourable values or results are shaded yellow, unfavourable or poor values are shaded grey. Any E value below 0.001 was considered a good score, any E value above 0.05 was considered poor. Sequences with mismatched cysteine residues in the aligned sequences were coloured grey, sequences with aligned cysteines were coloured yellow. Shared domains or motifs were considered a favourable result and so were coloured yellow accordingly. Any result that was at all ambiguous was left uncoloured. By allocating positive and negative values to yellow and grey results respectively, a rudimentary score was obtained for each sequence alignment. Each BLAST hit alignment was considered individually, and alignments with negative scores were deemed unlikely to represent homologous genes with conserved function.

After this final evaluation and filtering step, the list was reduced to a final list of 10 candidate genes (Table 4.3). Most of these genes are already well known to be associated with immune functions or are already established therapeutic targets with drugs already on the market. CD80, IL-1 and TNF are well established targets for drugs currently on the market for arthritis. CCR7 and CXCR2 are receptors for chemokines, small proteins responsible for controlling directed chemotaxis of cells, particularly immune cells. IL-10 is an anti-inflammatory cytokine. There were, however, four genes (denoted as TGOI-1, -2, -3, -4) which have no associated pre-existing targeted therapies, and in some cases little precedent for association with known immune cell functions. These four genes will be taken forward for further evaluation as potential novel therapeutic targets by consideration as potential new therapeutic targets by UCB scientists.

Most of the genes in this final list, or their interacting differentially expressed Treg genes, have been associated with various forms of human autoimmune disease (Figure 4.8 & Table 4.4). The evidence supporting these associations included literature

evidence for changes in expression or function in diseased clinical samples. A genetic association with autoimmune disease has been observed for some of these putative tolerogenic genes. These include high penetrance, Mendelian inherited genetic variants which are recorded in the OMIM (Online Mendelian Inheritance in Man) database. In addition, genome wide association studies (GWAS) have found single nucleotide polymorphisms (SNPs) in some of these genes that associate with auto-immune diseases (Table 4.4).

4.5 Identification of putative functional modules

A PPI network comprising 2614 nodes representing proteins, and 5925 edges representing binary interactions between proteins was generated in Cytoscape. Putative functional modules were identified using the Cytoscape plugin, JActiveModules. Functional modules are highly inter-connected sub-networks which are enriched for significantly differentially expressed genes.

The nodes representing the ten significant viral hits were used as seed nodes from which to initiate the search. When the default parameters, Search depth = 1, Max depth = 2, were applied, four modules with significant scores (Z >3) were returned (Table 4.5). The search depth was adjusted and further searches carried out so that most of the genes of interest were included within the final set of modules. Details of seven sets of functional modules containing the putative tolerogenic genes are shown in Table 4.5. These functional modules cannot be shown as they contain details and information on functions of the four genes of interest to UCB as potential target candidates. However, an example of modules returned as a result of an alternative search using the six precedented viral hits, minus the four genes of interest, is shown in Figure 4.9. The interactions have confidence scores associated with them according to the MIscore results returned from PSISCORE.

The Cytoscape plugin, BiNGO, was used to assess if the nodes in the functional modules were enriched for biological processes recorded in the GeneOntology (GO) database. Enriching predicted modules for GO terms describing biological processes

offered further information on potential biological functions of modules. Only the top ten functional annotations are shown (Table 4.5) as the lists of results were extensive, with hundreds of GO annotations being returned in some cases. As would be expected, most functional modules are enriched for molecules involved in immune cell functions (Table 4.5). Those modules without immune cell functions in their top ten GO annotations did include significantly scoring immunity-related GO terms which were outside the top ten results listed.

Table 4.1 *Reciprocal BLAST results*

Human gene	Viral accession	Virus	Virus protein description	E value	% ID of HSP	HSP length	Length of query participating in HSP	Number of identical amino acids	Number of conserved amino acids	% coverage of human protein
ANXA7	NP_050259.1	Human herpesvirus 6	DNA replication [Human herpesvirus 6]	0.088	28.8	80	77	23	37	15.8
CCR7	YP_073753.1	Human herpesvirus 7	envelope glycoprotein UL33 [Human herpesvirus 7]	3.00E-11	21.3	287	282	61	117	74.6
CD2	NP_040870.1	Human adenovirus F	membrane glycoprotein E3 CR1-beta [Human adenovirus F]	0.007	27.4	106	93	29	47	26.5
CD79A	YP_002213790.1	Human adenovirus B	membrane glycoprotein E3 CR1-gamma [Human adenovirus B]	0.27	28.1	96	94	27	39	41.6
CD80	YP_001129512.1	Human herpesvirus 4 type 2	BARF1 [Human herpesvirus 4 type 2]	2.00E-06	31.6	114	108	36	56	37.5
CD80	YP_401719.1	Human herpesvirus 4	BARF1 [Human herpesvirus 4]	2.00E-06	31.6	114	108	36	56	37.5
C3	NP_042150.1	Variola virus	hypothetical protein VARVgp106 [Variola virus]	6.3	29.3	82	82	24	37	4.9
CXCR2	YP_001129433.1	Human herpesvirus 8	ORF74 [Human herpesvirus 8]	8.00E-26	30.3	221	217	67	115	60.3
FAM46C	YP_002302228.1	Rotavirus A	VP3 [Rotavirus A]	0.15	25.2	127	117	32	58	29.9
GNA12	YP_232989.1	Vaccinia virus	virion core protein [Vaccinia virus]	0.72	23.0	122	105	28	58	27.6
GP160	YP_001129392.1	Human herpesvirus 8	ORF39 [Human herpesvirus 8]	0.009	21.8	170	160	37	67	47.3
IL10	YP_001129439.1	Human herpesvirus 4 type 2	BCRF1 [Human herpesvirus 4 type 2]	1.00E-54	73.2	142	142	104	108	79.8
IL10	YP_401634.1	Human herpesvirus 4	BCRF1 [Human herpesvirus 4]	1.00E-54	73.2	142	142	104	108	79.8
IL10	YP_081552.1	Human herpesvirus 5	interleukin-10 [Human herpesvirus 5]	3.00E-07	24.0	171	170	41	68	95.5
IL1R2	YP_233079.1	Vaccinia virus	IL-1-beta-inhibitor [Vaccinia virus]	3.00E-34	30.3	314	312	95	146	78.4
IL1R2	NP_042232.1	Variola virus	hypothetical protein VARVgp188 [Variola virus]	2.00E-08	24.7	308	299	76	122	75.1
IL1R2	NP_042929.1	Human herpesvirus 6	DNA packaging protein UL32 [Human herpesvirus 6]	0.069	26.0	123	111	32	54	27.9
ITGB1	YP_001911113.1	Whitewater Arroyo virus	glycoprotein G1+G2 precursor [Whitewater Arroyo virus]	0.008	39.0	59	58	23	34	7.3
ITGB1	YP_001649226.1	Bear Canyon virus	glycoprotein precursor [Bear Canyon virus]	0.12	35.6	59	58	21	31	7.3
LGALS3BP	NP_040510.1	Human adenovirus C	control protein E1B 19K [Human adenovirus C]	0.71	29.6	115	111	31	52	19.0
LRP1	NP_620108.1	Langat virus	polyprotein [Langat virus]	0.32	27.5	160	143	44	68	3.1
PLXB1	NP_040894.1	Human papillomavirus type 4	hypothetical protein HpV4gp6 [Human papillomavirus type 4]	0.067	24.8	157	156	39	64	7.3
PTPRF	NP_040299.1	Human papillomavirus type 6b	regulatory protein E2 [Human papillomavirus type 6b]	0.39	30.8	91	86	28	35	4.5
RUFY3	YP_081543.1	Human herpesvirus 5	tegument protein UL14 [Human herpesvirus 5]	0.064	29.1	79	74	23	42	15.8
RUFY3	YP_073811.1	Human herpesvirus 7	myristylated tegument protein [Human herpesvirus 7]	0.14	33.8	68	68	23	32	14.5
SG223	YP_068022.1	Human adenovirus E	encapsidation protein IVa2 [Human adenovirus E]	1.1	25.3	87	85	22	37	6.1
SG223	NP_040852.1	Human adenovirus F	encapsidation protein IVa2 [Human adenovirus F]	1.1	27.1	85	83	23	35	5.9
SORL	NP_043429.1	Human papillomavirus type 50	major capsid protein L1 [Human papillomavirus type 50]	0.59	29.7	74	66	22	32	3.0
TNFRSF1B	NP_042240.1	Variola virus	hypothetical protein VARVgp196 [Variola virus]	9.00E-36	39.9	183	176	73	95	38.2
TNFRSF1B	YP_233061.1	Vaccinia virus	secreted TNF-receptor-like protein [Vaccinia virus]	4.00E-07	40.0	55	55	22	30	11.9
UBP13	NP_042907.2	Human herpesvirus 6	protein U15 [Human herpesvirus 6]	0.098	25.9	81	81	21	39	9.4
UTRO	NP_044592.1	Respiratory syncytial virus	Phosphoprotein (P) [Respiratory syncytial virus]	0.71	28.2	85	85	24	37	2.5
UTRO	YP_081540.1	Human herpesvirus 5	DNA packaging tegument protein UL17 [Human herpesvirus 5]	3.5	33.3	66	61	22	32	1.8
VLDLR	NP_941979.1	Uukuniemi virus	membrane glycoprotein polyprotein [Uukuniemi virus]	0.034	22.8	171	165	39	67	18.9
TGOI-1	Virus a ID	Virus a	Protein [Virus a]	5.00E-13	26.8	220	212	59	101	4.8
TGOI-1	Virus b ID	Virus b	Protein [Virus b]	3.00E-12	22.5	338	303	76	147	6.9
TGOI-2	Virus c ID	Virus c	Protein [Virus c]	1.00E-21	56.9	72	72	41	55	23.8
TGOI-3	Virus d ID	Virus d	Protein [Virus d]	0.001	26.6	154	148	41	62	29.4
TGOI-4	Virus f ID	Virus e	Protein [Virus e]	5.00E-15	41.0	78	78	32	43	27.6
TGOI-4	Virus g ID	Virus f	Protein [Virus f]	9.00E-06	35.5	62	62	22	29	21.9

Figure 4.4 Proteins upregulated in Treg cells with homology to viral proteins

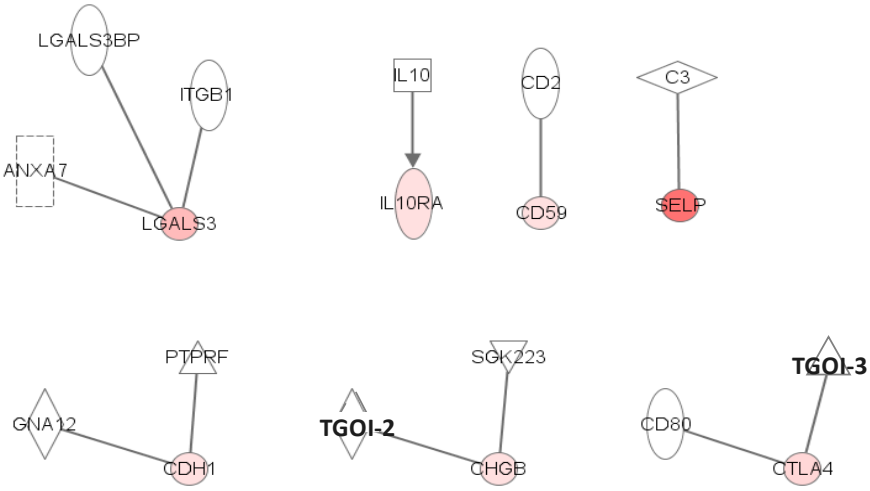
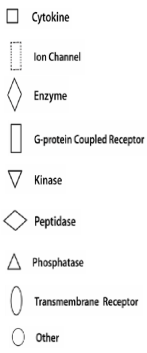


Figure 4.5 Proteins with homology to viral sequences that interact with upregulated Treg proteins

Red = Upregulated in Treg cells

White = interacting protein, viral homologue

Legend for figures 4.4 & 4.5



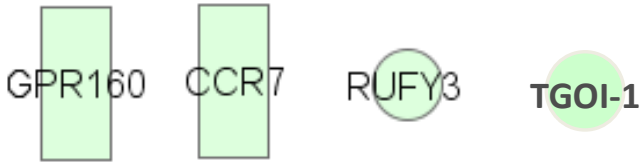


Figure 4.6 Proteins downregulated in Treg cells with homology to viral proteins

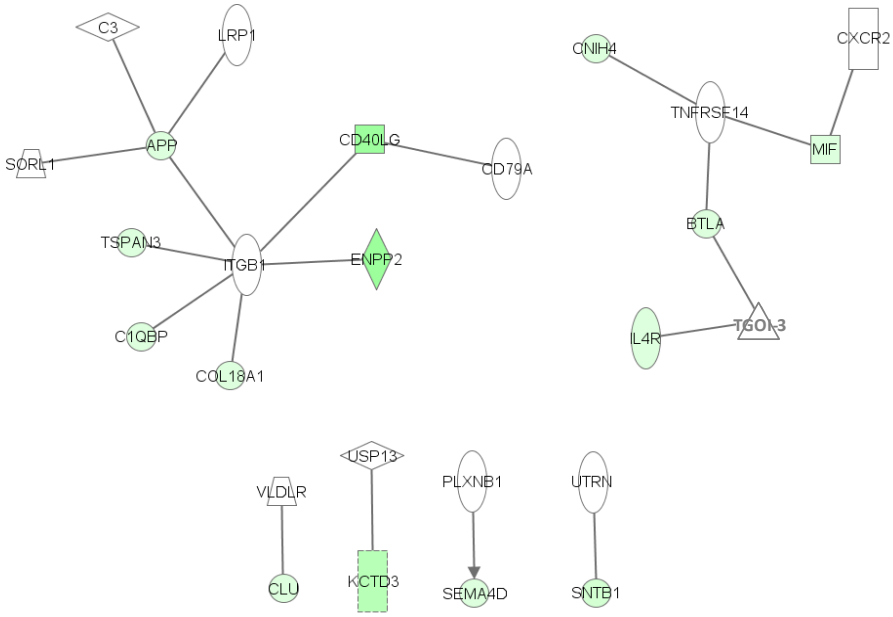


Figure 4.7 Proteins with homology to viral sequences that interact with downregulated Treg proteins

Green = Downregulated in Treg cells

White = interacting protein, viral homologue

Legend for figures 4.6 & 4.7:

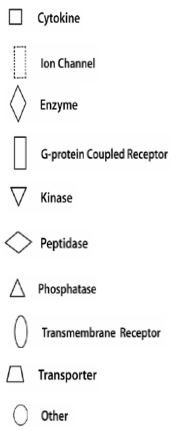


Table 4.2 *Evaluation of aligned sequences*

Human gene	Virus hit - protein description	E value	Matching cysteines in hsp?	Superfamily motifs - virus	Number yellow	Number grey	score (y-g)	Valid hit?
ANXA7	DNA replication [Human herpesvirus 6]	0.088	1 yes 1 no	none	0	2	-2	NO
CCR7	envelope glycoprotein UL33 [Human herpesvirus 7]	3.00E-11	YES	Family A G protein-coupled receptor-like	3	0	3	YES
CD2	membrane glycoprotein E3 CR1-beta [Human adenovirus F]	0.007	NO	Immunoglobulin	1	1	0	NO
CD79A	membrane glycoprotein E3 CR1-gamma [Human adenovirus B]	0.27	NO	Immunoglobulin	0	2	-2	NO
CD80	BARF1 [Human herpesvirus 4 type 2]	2.00E-06	YES	Immunoglobulin	3	0	3	YES
	BARF1 [Human herpesvirus 4]	2.00E-06	YES	Immunoglobulin	3	0	3	YES
C3	hypothetical protein VARVgp106 [Variola virus]	6.3	NO	none	0	3	-3	NO
CXCR2	ORF74 [Human herpesvirus 8]	8.00E-26	YES	Family A G protein-coupled receptor-like	3	0	3	YES
FAM46C	VP3 [Rotavirus A]	0.15	NO	none	0	2	-2	NO
GNA12	virion core protein [Vaccinia virus]	0.72	NO	none	0	3	-3	NO
GP160	ORF39 [Human herpesvirus 8]	0.009	NO	none	0	2	-2	NO
IL10	BCRF1 [Human herpesvirus 4 type 2]	1.00E-54	YES	4-helical cytokines (Interferon/IL-10 family)	3	0	3	YES
	BCRF1 [Human herpesvirus 4]	1.00E-54	YES	4-helical cytokines (Interferon/IL-10 family)	3	0	3	YES
	interleukin-10 [Human herpesvirus 5]	3.00E-07	YES	4-helical cytokines (Interferon/IL-10 family)	3	0	3	YES
IL1R2	IL-1-beta-inhibitor [Vaccinia virus]	3.00E-34	YES	Immunoglobulin	3	0	3	YES
	hypothetical protein VARVgp188 [Variola virus]	2.00E-08	YES	Immunoglobulin	3	0	3	YES
	DNA packaging protein UL32 [Human herpesvirus 6]	0.069	NO	none	0	3	-3	NO
ITGB1	glycoprotein G1+G2 precursor [Whitewater Arroyo virus]	0.008	NO	none	0	2	-2	NO
	glycoprotein precursor [Bear Canyon virus]	0.12	NO	none	0	3	-3	NO
LGALS3BP	control protein E1B 19K [Human adenovirus C]	0.71	No cys in hsp sequence	none	0	2	-2	NO
LRP1	polyprotein [Langat virus]	0.32	NO	none	0	3	-3	NO
PLXB1	hypothetical protein HpV4gp6 [Human papillomavirus type 4]	0.067	NO		0	3	-3	NO
PTPRF	regulatory protein E2 [Human papillomavirus type 6b]	0.39	NO	none	0	3	-3	NO
RUFY3	tegument protein UL14 [Human herpesvirus 5]	0.064	NO	none	0	3	-3	NO
	myristylated tegument protein [Human herpesvirus 7]	0.14	NO	none	0	3	-3	NO
SG223	encapsidation protein IVa2 [Human adenovirus E]	1.1	1 yes, 1 no	P-loop containing nucleoside triphosphate hydrolases	0	2	-2	NO
	encapsidation protein IVa2 [Human adenovirus F]	1.1	1 yes, 1 no	P-loop containing nucleoside triphosphate hydrolases	0	2	-2	NO
SORL	major capsid protein L1 [Human papillomavirus type 50]	0.59	NO	none	0	3	-3	NO
TNFRSF1B	hypothetical protein VARVgp196 [Variola virus]	9.00E-36	YES	TNF receptor-like	3	0	3	YES
	secreted TNF-receptor-like protein [Vaccinia virus]	4.00E-07	YES	TNF receptor-like	3	0	3	YES
UBP13	protein U15 [Human herpesvirus 6]	0.098	No cys in hsp sequence	none	0	2	-2	NO
UTRO	Phosphoprotein (P) [Respiratory syncytial virus]	0.71	NO	none	0	3	-3	NO
	DNA packaging tegument protein UL17 [Human herpesvirus 5]	3.5	No cys in hsp sequence	none	0	2	-2	NO
VLDLR	membrane glycoprotein polyprotein [Uukuniemi virus]	0.034	4 YES; 16 NO	none	0	2	-2	NO
TGOI-1	Virus a	5.00E-13	YES	Matched domain	3	0	3	YES
	Virus b	3.00E-12	YES	Matched domain	3	0	3	YES
TGOI-2	Virus c	1.00E-21	No cys in hsp sequence	Matched domain	2	0	2	YES
TGOI-3	Virus d	0.001	1 yes 1 no	Matched domain	1	0	1	YES?
TGOI-4	Virus e	5.00E-15	YES	Matched domain	3	0	3	YES
	Virus f	9.00E-06	YES	Matched domain	3	0	3	YES

Yellow = favourable value, Grey = poor value, White = Inconclusive/not applicable

E value < 0.001 = yellow, E value > 0.05 =Grey.

Table 4.3 *Final list of putative tolerogenic genes*

Gene	Protein name	Cellular location
CCR7	Chemokine (C-C motif) receptor 7	PM
CD80	CD80 molecule	PM
CXCR2	Chemokine (C-X-C motif) receptor 2	PM
IL10	Interleukin 10	EC
IL1R2	Interleukin 1 receptor, type II	PM
TNFRSF1B	Tumor necrosis factor receptor superfamily, member 1B	PM
TGOI-1	-	PM
TGOI-2	-	PM
TGOI-3	-	PM
TGOI-4	-	PM

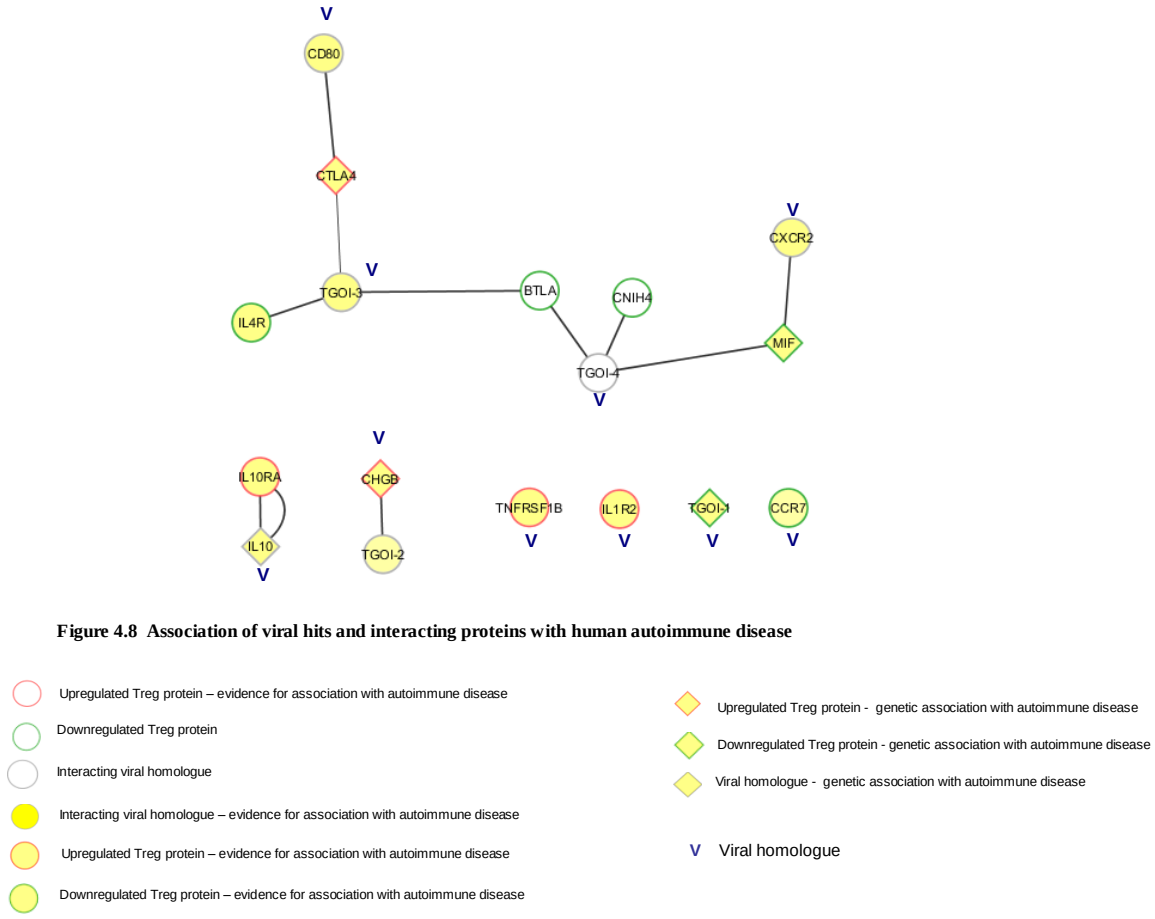


Figure 4.8 Association of viral hits and interacting proteins with human autoimmune disease

Table 4.4 ***Putative tolerogenic genes and their interactors – association with autoimmune diseases***

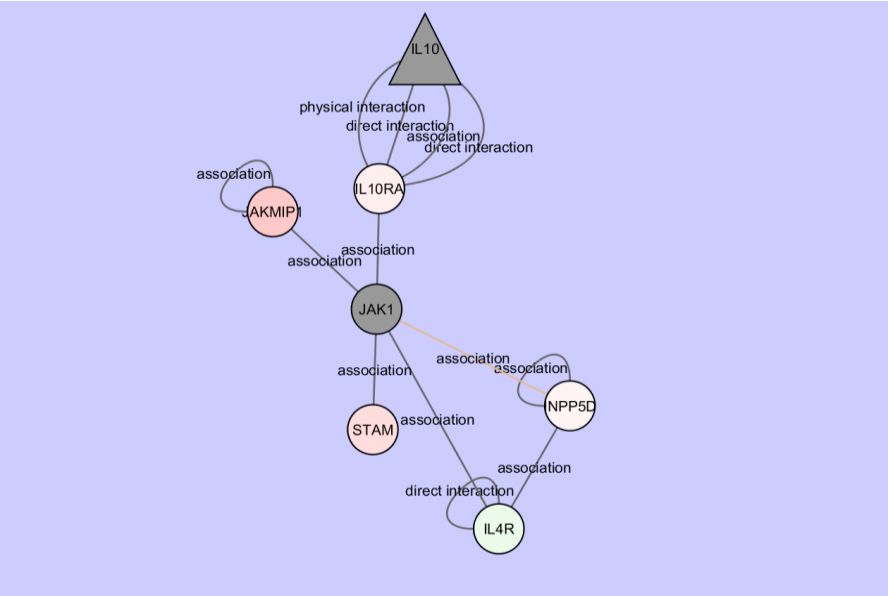
	Rheumatoid arthritis	Type 1 diabetes	Systemic lupus erythematosus	Pemphigus/phemigoid	Grave's disease	Hashimoto thyroiditis	Autoimmune haemolytic anaemia	Coeliac disease	Sjogren's syndrome	Evidence for genetic association with disease?
Upregulated Treg genes										
IL10RA	✓									SNP, RA. dBSNP: rs11032362
CHGB	✓									SNP, RA. dBSNP: rs236151
IL1R2 (v)	✓									
TNFRSF1B (v)	✓									
CTLA4	✓	✓		✓	✓	✓		✓		IDDM. OMIM: 601388
										SNP, IDDM. dBSNP: rs3087243
										Graves Disease. OMIM: 275000
										Hashimoto's thyroiditis. OMIM: 140300
										Coeliac disease 3. OMIM: 609755
Interacting viral homologues										
CD80 (v)	✓		✓							
IL10 (v)	✓	✓	✓							RA. OMIM: 180300
TGOI-3 (v)							✓			
Downregulated Treg genes										
CCR7 (v)									✓	
MIF	✓									Juvenile RA: OMIM: 604302
IL4R	✓								✓	
TGOI-1 (v)		✓								IDDM SNP
Interacting viral homologues										
CXCR2 (v)	✓									

Information obtained from IPA and NextBio

Table 4.5 ***Functional modules generated with JActiveModules.***
Only the top 10 biological processes for each module are shown

Max search depth	Nodes	Edges	Score	Number of proteins from module in GO biological process	Top 10 biological processes - GO ontology
2	80	261	11.4	18	*regulation of cell death*
				17	*regulation of apoptosis*
				17	*regulation of programmed cell death*
				18	*cellular component biogenesis*
				3	*regulation of necrotic cell death*
				7	*regulation of cellular response to stress*
				7	*regulation of protein transport*
				6	*regulation of intracellular transport*
				7	*regulation of establishment of protein localization*
				12	*regulation of phosphorus metabolic process*
2	27	62	9.08	3	*activation of pro-apoptotic gene products*
				2	*primary microRNA processing*
				4	*developmental growth*
				2	*pericardium development*
				2	*SMAD protein complex assembly*
				2	*paraxial mesoderm morphogenesis*
				2	*embryonic foregut morphogenesis*
				2	*foregut morphogenesis*
				2	*positive regulation of cell morphogenesis involved in differentiation*
				2	*positive regulation of epithelial to mesenchymal transition*
2	17	37	4.41	6	*regulation of cell activation*
				4	*negative regulation of lymphocyte activation*
				9	*immune system process*
				4	*negative regulation of leukocyte activation*
				4	*negative regulation of cell activation*
				5	*negative regulation of cell activation*
				5	*regulation of lymphocyte activation*
				4	*regulation of leukocyte activation*
				3	*negative regulation of immune system process*
					negative regulation of lymphocyte proliferation
2	6	9	7.65	2	*negative regulation of alpha-beta T cell proliferation*
				3	*regulation of lymphocyte proliferation*
				3	*regulation of mononuclear cell proliferation*
				3	*regulation of leukocyte proliferation*
				2	*negative regulation of alpha-beta T cell activation*
				2	*regulation of alpha-beta T cell proliferation*
				3	*regulation of lymphocyte activation*
				3	*regulation of leukocyte activation*
				3	*regulation of cell activation*
				2	*negative regulation of T cell proliferation*
3	119	335	17.98	31	*regulation of apoptosis*
				31	*regulation of programmed cell death*
				31	*regulation of cell death*
				47	*positive regulation of biological process*
				43	*positive regulation of cellular process*
				20	*positive regulation of apoptosis*
				20	*positive regulation of programmed cell death*
				78	*regulation of cellular process*
				20	*positive regulation of cell death*
				17	*induction of apoptosis*
3	31	78	5.56	8	*regulation of cell activation*
				21	*regulation of response to stimulus*
				21	*regulation of immune system process*
				21	*positive regulation of cell activation*
				7	*signal transduction*
				6	*positive regulation of biological process*
				6	*regulation of lymphocyte activation*
				6	*regulation of cellular process*
				7	*positive regulation of immune system process*
				8	*immune system process*
4	93	259	15.66	8	*negative regulation of cell activation*
				10	*regulation of apoptosis*
				9	*regulation of programmed cell death*
				6	*regulation of cell death*
				15	*negative regulation of lymphocyte activation*
				16	*negative regulation of leukocyte proliferation*
				6	*negative regulation of lymphocyte proliferation*
				25	*negative regulation of mononuclear cell proliferation*
				7	*negative regulation of leukocyte activation*
				11	*regulation of lymphocyte proliferation*

(a)



(b)

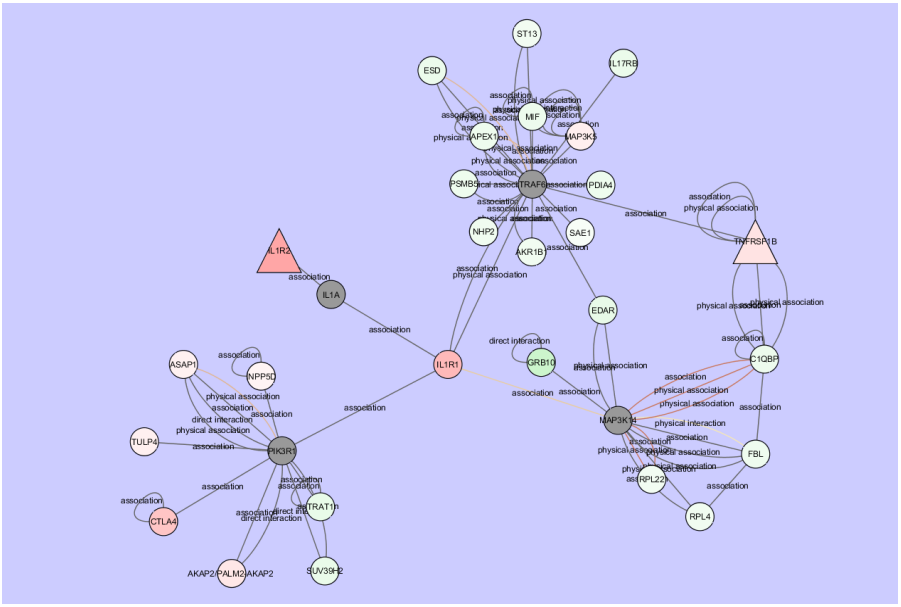


Figure 4.9 Examples of functional modules returned using JActiveModules.
Six of the final list of tolerogenic genes were used as seed nodes, and a search was performed using parameters set as depth=2, max depth from nodes=3. The network in (a) scored 12.71, the network in (b) scored 4.01. Red nodes = increased expression in Treg cells, Green nodes = decreased expression in Treg cells. Triangular nodes = proteins with viral homologues. Intensity of colour is proportional to gene expression values. Edges are coloured in proportion to confidence scores for the interactions [see legend in (a)] Grey edges= no score was returned from PSISCORE.

5 Discussion

5.1 Generation of a consensus Treg gene signature

A consensus Treg gene signature was generated in this study following a meta-analysis to compare multiple public Treg vs non-Treg gene expression studies. Gene expression data is inherently noisy due to sample heterogeneity, probe promiscuity, and stochasticity of biochemical events such as promoter binding, gene transcription etc. This fundamental biological noise is further compounded by variability due to different experimental laboratories using different expression platforms. Public gene expression repositories contain thousands of independent studies often with numerous studies investigating the same phenomenon. The meta-analysis approach, propounded here, exploits the availability of public experimental "replicates" to mitigate the effects of errors that occur within a single experiment on a single platform, since the reliability of an observation is weighted by its consistency across multiple studies and platforms (Kuperschmidt *et al.*, 2010).

A true consensus gene signature for Treg cells will naturally only be generated if the studies included in the meta-analysis are of good quality and do actually reflect a true comparison of Treg cells with non-Treg cells. There is currently no single, definitive marker for the identification of Treg cells. In most of the published studies considered for this analysis, Treg cells were isolated from peripheral blood by selecting for T cells positive for CD4 and CD25 expression, and also FOXP3 in some cases. However, CD25 is expressed on around a quarter of all $CD4^+$ T cells, and it is thought that only the very highest CD25 expressers are Treg cells (Shevach, 2006). Also, in contrast to the mouse, FOXP3 expression in humans may not be completely confined to $CD25^+CD4^+$ cells. It has been observed that human $CD4^+$ $CD25^-$ T cells may express FOXP3 mRNA upon T cell receptor (TCR) stimulation, although the expression levels are generally much lower and more transient than in Treg cells (Walker *et al.*, 2003; Morgan *et al.*, 2005). Some studies also distinguished between cells expressing high and low levels of CD127 (IL-7 receptor α chain), another marker suggested to discriminate between Treg and conventional $CD4^+$ T cells; but it has been reported that

upon activation, most $CD4^+$ cells downregulate CD127 (Corthay *et al.*, 2009). Datasets were therefore selected for inclusion in the analysis based upon the increased expression of a number of pre-determined key Treg-associated genes in combination This gene combination was decided upon in consultation with experts in the field at UCB. Individually these genes may not be definitive markers for Treg cells, but together they should provide good evidence for a Treg phenotype.

A similar study was previously carried out at UCB, with the aim of identifying tolerogenic genes in the mouse. In that study, the meta-analysis was performed across five datasets, and differential expression of 605 genes was observed in four out of the five studies. This was double the number of consensus genes found in the current study where 303 differentially expressed genes were observed in four of six datasets. This is most likely due to the inherent variability often seen in biological samples taken from human subjects and may be further compounded by the difficulty of precisely isolating Treg immune cells. It might be expected that highly inbred mouse strains, housed in controlled laboratory conditions will represent a far more homogenous population, displaying a greater degree of similarity in their gene expression profiles. Therefore it would be expected that more genes would be included in a consensus signature across multiple mouse datasets, while less genes would form a consensus across human studies with greater inter-dataset variability.

5.2 Expansion of the Treg gene signature to include interacting proteins

In the present study, proteins interacting with Treg gene products were identified using IPA software. A total of 516 Treg proteins and direct interactors were found located in the plasma membrane, extracellular space or in unknown locations. In the previous mouse study, interacting proteins were identified using a different software tool called Pathway Studio, with which 1176 mouse Treg and interacting proteins were identified. The magnitude of the expansion of the gene list is slightly higher in the mouse study than in the human study; 94% amplification in the mouse compared to 70% in the human. This difference could be attributed to the different software tools utilized. The

core platform technology used by Pathway Studio is their "MedScan Technology", software that uses text-mining algorithms to automatically extract data from literature sources in an unsupervised manner. In contrast, all information in the IPA database is manually reviewed and curated. As well as information from published literature, it also includes information from a range of third party sources and databases. These differing approaches to attaining PPI data could result in the IPA software applying more rigorous filters to its data acquisition than Pathway Studio, resulting in more interactions being excluded from the analysis. The differences may also conceivably be due to there being more hub proteins in the mouse consensus Treg gene list, resulting in a greater number of interactions and neighbouring proteins.

The rationale behind this study was to identify putative tolerogenic genes that could become candidate targets for novel therapeutics. To be useful pharmaceutically, such a target needs to be druggable, i.e. it needs to be accessible to putative therapeutic agents. In this case, the prospective therapeutic agent will be a monoclonal antibody. Monoclonal antibodies are attractive for pharmaceutical companies as they are considered to be easier to develop than small molecule drugs (e.g. they have less off-target effects). The estimated time and cost to bring a monoclonal antibody to the stage where it is ready for clinical testing is significantly less than that needed for a traditional small molecule drug (Ezzell, 2001). In the present study, putative targets were selected based upon their localization at the cell surface or in the extracellular space, making them accessible to antibodies. IPA provides information on cellular localizations of molecules, based upon GO cellular compartment annotations and it was this information that was used as our "druggability" filter.

5.3 Search for viral homologues

Many viruses are known to co-opt genes for host proteins, allowing the virus to manipulate detection and elimination by the host immune system. This is especially true for viruses that establish persistent infections and reside in the host for long periods of time. Genes co-opted by a virus may steer us towards powerful points of intervention in an immune response. Treg cells are central to the maintenance of

tolerance, so some genes that are specifically differentially expressed in Treg cells may be important for tolerogenic mechanisms. The hypothesis behind the approach taken in the present study is that some viruses may have co-opted genes that can induce tolerance. The presence of differentially expressed Treg genes in viral genomes was therefore used as a stringent filter to select putative tolerogenic genes from our network expanded consensus Treg gene signature. Interestingly, of the six genes in the final list of ten that can be discussed here, four of them (CCR7, CD80, CXCR2 and IL-10) aligned with protein sequences present in herpes viruses, viruses known to establish persistent infections. Indeed, herpes viruses have the ability to persist as a lifelong latent infection due to their successful coexistence with the host, facilitated by numerous mechanisms acquired for modulating the host immune system.

Human IL-10, an anti-inflammatory cytokine, that aligned with BCRF1 (Bam HI C fragment rightward reading frame) protein (also known as viral IL-10 homologue) of human herpes virus 4 (Epstein Barr virus [EBV]) and IL-10 protein of herpes virus 5 (Human Cytomegalovirus [HCMV]). It has previously been established that herpes viruses, such as HCMV and EBV can express their own viral cytokines, or 'virokines'. The IL10 virokine is biologically active and retains the immunosuppressive activities of its host counterpart (Hsu *et al.*, 1990; Spencer *et al.*, 2002). Human IL1R2 and TNFRSF1B aligned with proteins present in the poxviruses, vaccinia and variola. These IL1R2 (IL-1 receptor) and TNFR (TNF receptor) viral homologues are soluble forms of the receptors and have been shown to bind to IL-1 and TNFα, respectively, acting as a 'sink' to reduce the inflammatory effects of these cytokines (Haig, 1998).

The above-mentioned studies on viral homologues of cytokines and cytokine receptors demonstrated that although they shared only 20-30% amino acid identity with their human counterparts, immunosuppressive functions were retained. This raises the issue of statistical versus biological significance when evaluating the results of BLAST search alignments. The E value associated with a BLAST result is a measure of the statistical significance of the alignment, related to the probability of the alignment occurring by chance in a database of a particular size. The E value is related to the lengths of the query sequence and search space (i.e. database size) and identity between the query and target sequences. E-values above 0.01 are generally considered to be

dubious. Statistics do not, however, necessarily reflect biological significance, which was in this context equates to protein function. It is possible that viruses may only need to incorporate a small portion of a human gene to retain its vital immuno-modulatory functions. Co-opted viral genes may also have become modified and evolved over time to continually maintain optimal benefits for virus survival and persistence (Griffin *et al.*, 2010). These factors could result in low sequence identity and correspondingly high E values, while retaining biologically significant functions. Increasing the E value threshold when setting BLAST parameters may result in statistically insignificant alignments, while revealing biologically significant results. Therefore, rather than solely considering the statistical outputs of the BLAST searches, biological significance of the BLAST results was also evaluated by manual inspection of the alignments between query and hit sequences. Each alignment was inspected to adjudge the conservation of cysteine residues in the sequences. Cysteine is an amino acid important in maintaining tertiary structure and function of proteins through its participation in the formation of disulphide bridges. Functional similarity between viral and human proteins was further examined by searching for shared functional motifs and protein family domains. This approach reduced the candidate list of tolerance gene from 27 to 10.

Most of the genes in the final list of ten are already known to be associated with immune functions or are established therapeutic targets with drugs already on the market. CCR7 and CXCR2 are receptors for chemokines, small proteins responsible for controlling directed chemotaxis or movement of cells, particularly immune cells. IL-10 is an anti-inflammatory cytokine. CD80, the IL-1 receptor (IL-1R) and TNFα receptor (TNFR) are well established targets for drugs currently on the market for the autoimmune disease, rheumatoid arthritis (RA).

IL-1 is a pro-inflammatory cytokine that drives joint inflammation in RA. The drug, Anakinra (marketed as 'Kineret', Amgen Inc.) is a recombinant IL-1R antagonist which competes with IL-1 for binding with the naturally occurring form of the IL1R, thereby blocking biologic activities of IL-1, including inflammation and bone resorption, and cartilage degradation associated with RA (Furst, 2004).

TNFα is another pro-inflammatory cytokine. Elevated levels of TNFα are found in tissues and fluids of patients with autoimmune disease such as rheumatoid arthritis, psoriatic arthritis, ankylosing spondylitis, and plaque psoriasis. A number of biologic therapies targeting TNF are currently on the market (Kuek *et al.*, 2007). Etanercept (Enbrel, marketed by Amgen, Pfizer, & Wyeth) is an anti-TNF drug, consisting of an engineered TNFR2 dimeric fusion protein. It acts as a decoy receptor, preventing TNF from interacting with natural TNFRs. There are various anti-TNF monoclonal antibody therapeutics on the market, such as Certolizumab (Cimzia, UCB), Adalimumab (Humira, Abbott) and Infliximab (Remocade, Centocor & Merck).

CD80 is expressed on antigen presenting cells and provides a co-stimulatory signal necessary for T cell activation. It is the ligand for CD28, which is expressed on T cells, and CTLA-4, which is expressed by Treg cells where it causes attenuation of the co-stimulatory activation signal. CTLA-4 coupled to the Fc domain of an immunoglobulin G1 molecule blocks this CD80-mediated co-stimulatory pathway, and is marketed as Abatacept, or Orencia™ (Bristol Myers Squibb). It is used for the treatment of rheumatoid arthritis (Vincenti & Luggen, 2007). Another blocker of this pathway, the anti-CD80 monoclonal antibody, Galiximab (Biogen Idec), has shown efficacy in clinical trials for psoriasis (Gottlieb *et al.*, 2004).

The inclusion of a number of known therapeutic targets for autoimmune diseases in the final list of putative tolerogenic genes serves to support the fundamental hypothesis underpinning this approach; namely the likely co-option of human immuno-modulatory genes by viruses, and validates the method. Although those genes may not provide new insights or offer novel targets to be exploited, their presence in the final set of results indicates that the methodology enables identification of genes which may be involved in the dysregulation of self-tolerance and that can be successfully targeted therapeutically. Their presence alongside the four genes of interest for further consideration puts these four genes in good company, and strengthens the case for their potential as new targets.

5.4 Identification of putative functional modules

The biological rationale supporting the involvement of the putative targets in tolerance was further explored by generating focused sub-networks that are active in Treg vs non-Treg cells and are seeded on the putative target proteins. This was achieved by generating a Treg-specific PPI network and obtaining confidence scores for the interactions before identifying functional modules within the network.

The Treg PPI interaction network was generated by consolidating PPI data from a number of different PPI data sources. This was necessary because it has been shown that there are low levels of agreement between databases when curating PPI data from the same publications, and that there is often little overlap between databases (Turinsky *et al.*, 2011). The PSICQUIC interface, a federated search resource, was used to retrieve interaction data from the different interaction data sources. Data was exported from each selected data source for subsequent generation of confidence scores.

Since completing this study, following a discussion with the developers of PSICQUIC, the web interface has been changed, so that when opting to cluster results (remove redundant binary interactions where the same interaction is found in more than one database), individual data sources can be selected for clustering and not all the returned results have to be clustered as previously (see 'Methods', Section 3.5.1). This could simplify the workflow, as it would allow redundant interactions to be removed and would also standardize the node IDs, as the interacting proteins in the clustered results output are all returned with UniProt IDs.

Similarly to gene expression data, the quality of public PPI data can vary widely potentially limiting its utility. There are many different experimental techniques for determining molecular interactions, producing results of varying reliability, from high-throughput yeast 2-hybrid studies through to small-scale single protein studies, as well as methods utilizing computational predictions. Some of these experimental techniques are considered more likely to give rise to false positive results than others, particularly some high-throughput methods such as yeast 2-hybrid (Sprinzak *et al.*, 2003) and affinity purification coupled with tandem mass spectroscopy (TAP-MS) [Lavallee-Adam *et al.*, 2011).

The PSI confidence scoring system (PSISCORE) was developed to provide a means for assessing the quality and reliability of molecular interaction data (Aranda *et al.*, 2011). The MIscore method takes into account publications supporting an interaction and the types and number of experimental techniques used for detecting an interaction. Different detection methods are assigned scores for a particular interaction type. Experimental techniques considered to be less reliable for determining a particular interaction type will score poorly. PSISCORE and the MIscore method were used to associate confidence metrics with each edge in the Treg PPI network, enabling the network to be flexibly calibrated for accuracy as required. For example, low confidence associations could be maintained if the emphasis is on discovery whilst only high-quality interactions preserved if the requirement is to interpret a finding in the biological context of canonical interactions.

Functional modules or subnetworks can be identified as highly connected network regions which show significant changes in gene expression (Cline *et al.*, 2007). The identification of putative functional modules was made possible by the integration of Treg gene expression data with the Treg PPI data. Functional modules are enriched for differentially expressed genes but may contain genes that were not in the consensus Treg gene signature that are required to connect other differentially expressed genes. In this respect, the results differ to those that would be obtained using clustering techniques for gene expression studies that have not been integrated with PPI networks. Thus, integration of expression and PPI data may provide improved information on biological mechanisms and processes. It does this by enabling the identification of a gene that is not significantly differentially expressed, but may occupy an important position in the topology of a PPI network, locally enriched for differentially expressed interactors.

The putative functional modules identified in this study contained most of the ten putative tolerogenic genes in the final list, offering richer insight into their possible involvement in tolerance via the biological processes represented by the proteins and interactions evident in the identified active sub-networks. Enriching predicted modules for GO terms describing biological processes offered further information on potential

biological functions of modules, and provides a means of validating module predictions (Cline *et al.*, 2007).

These putative functional modules along with their associated GO biological process annotations, provided highly focused distillations of public omic datasets that are amenable for further evaluation in collaboration with UCB immunologists.

6 Conclusions & future work

6.1 Conclusions

The aim of this study was to identify putative tolerogenic genes for consideration as targets for the development of novel therapeutics for the treatment of immunological disorders. In this respect, the work can be considered to have been successful, in that four genes were identified that are now under investigation as potential new targets by UCB.

The methodology used to reveal these potential new targets has been validated to a large extent by the inclusion alongside the potentially novel targets, of known targets for autoimmune diseases.

The four candidate genes selected for further evaluation as potential targets have a number of associated pros and cons. Two of them are not readily associated with immunity or immune cell function, so could present truly novel targets that perhaps have never before been considered. However, the downside of this is that with so little being known about their biology and relevance to immunity, a lot of further scientific validation will be required, particularly in the form of 'wet' lab work.

More is known about the roles of the other two genes in immune cell functions. However, this advantage carries the accompanying potential downside that the IP (intellectual property) space may be more crowded. One of these known immuno-modulatory genes was a more ambiguous, borderline viral hit. The E value (0.001) for its BLAST alignment with the viral protein was at the boundary for what is statistically considered a good or poor alignment, and the alignment of cysteine residues in the human and viral proteins was inconclusive. It did share an immunoglobulin domain with its putative viral homologue, although the immunoglobulin domain is a broad structural domain classification and is seen quite frequently in viral genomes. The second of the genes with known immuno-modulatory function had only one piece of evidence supporting its interaction with an upregulated Treg gene, observed in a high

throughput yeast-2 hybrid screen. Its interaction with a differentially expressed Treg gene is therefore not well defined, so again, further investigation and validation will be required for this gene before it can be considered for further progression.

The study has now reached the stage where biologists and laboratory scientists can take on the further investigations required for possible progression of these genes as new targets.

To conclude, current immunosuppressive therapies are unsatisfactory for a variety of reasons. This work has successfully provided a novel computational approach for identifying putative tolerogenic genes by using viral genomes as a very stringent filtering mechanism. These genes that may control tolerance mechanisms could potentially represent targets for a new class of therapeutic that could induce tolerance and perhaps achieve the ultimate goal of long term remission of autoimmune disease.

6.2 Future work

There are a few changes and additional investigations that could potentially improve the workflow and provide additional insight into the results. Potential further application of the approaches used in the current study is also discussed in this section:

- Methods to improve the evaluation of the statistical results of the BLAST alignments could be further investigated. All human proteins known to have been co-opted by viruses could be BLASTed against viral genomes, and the resulting E values examined and evaluated more thoroughly. This could give a more precise idea of the range of E values commonly associated with co-opted proteins, information which could be applied to allow a more informed evaluation of unknown BLAST viral hits.
- It may be possible to streamline the workflow involved in generating a PPI network with associated confidence scores. As mentioned in the discussion, (Section 5), the web interface for PSICQUIC has recently been changed to improve clustering methods within the application. This could simplify the

workflow and standardize the node IDs. Sample files have been sent to the developers of PSISCORE and they are currently investigating the reasons for many binary interactions being returned with no associated confidence scores. It can be anticipated that these tools will continually improve, which can only enhance their application to studies of PPI networks, such as ours, and assessment of data quality.

- It would be interesting to experiment with other applications for analysing features of biological networks. There are many more plugins for Cytoscape for the analysis of networks and functional enrichment. It would be useful to examine these in more detail for possible application to our data for gaining further insights.

- In a similar manner to viruses, many parasitic and commensal organisms have the ability to modify host immunity. Helminths are known to produce molecules that mimic or modify molecules of the host immune system, which underpins their persistence. In countries endemic for parasitic helminth infections, autoimmune and allergic diseases remain relatively rare, so it has been hypothesized that helminths may protect against the development of autoimmunity and allergy (Harnett & Harnett, 2008). Tick saliva is known to contain a repertoire of components that have anti-haemostatic, anti-inflammatory and immunomodulatory effects which aids their ability to obtain a blood meal from the host (Valenzuela, 2004). Commensal bacteria have been shown to have a role in modulating mucosal immune responses (Forsythe & Bienenstock, 2010). As an extension of the present study, the genomes of parasitic organisms or commensal bacteria could be utilized in an analogous manner to viral genomes to look for putative modulators of the immune system that may provide intervention points in immune disorders.

REFERENCES

1. Alberghina, L., Hofer, T and Vanoni, M. (2009). Molecular networks and system-level properties. Journal of Biotechnology, 144: 224-233.

2. Albert, R. (2005). Scale-free networks in cell biology. Journal of Cell Science, 118: 4947-4957.

3. Aranda, B., Blankenburg H., Kerrien S., Brinkman F.S., Ceol A., Chautard E, Dana J.M., De Las Rivas J., Dumousseau M., Galeota E., Gaulton A., Goll J., Hancock R.E., Isserlin R., Jimenez R.C., Kerssemakers J., Khadake J., Lynn D.J., Michaut M., O'Kelly G., Ono K., Orchard S., Prieto C., Razick S., Rigina O., Salwinski L., Simonovic M., Velankar S., Winter A., Wu G., Bader G.D., Cesareni G., Donaldson I.M., Eisenberg D., Kleywegt G.J., Overington J., Ricard-Blum S., Tyers M., Albrecht M., Hermjakob H. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nature Methods 8: 528-529.

4. Asano, M., Toda, M. Sakaguchi, N, and Sakaguchi, S. (1996). Autoimmune disease as a consequence of developmental abnormality of a T cell subpopulation. Journal of Experimental Medicine, 184: 387–396.

5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25: 25-29.

6. Arrell, D.K. and Terzic, A. (2010). Network systems biology for drug discovery. Clinical Pharmacology and Therapeutics, 88: 120-125.

7. Battaglia M., Stabilini A., Migliavacca B., Horejs-Hoeck J., Kaupper T., Roncarolo M.G. (2006). Rapamycin promotes expansion of functional CD4+CD25+FOXP3+ regulatory T cells of both healthy subjects and type 1 diabetic patients. Journal of Immunology, 177:8338-47.

8. Bopp T., Becker C., Klein M., Klein-Hessling S., Palmetshofer A., Serfling E., Heib V, Becker M, Kubach J, Schmitt S, Stoll S, Schild H, Staege M.S., Stassen M, Jonuleit H, Schmitt E. (2007). Cyclic adenosine monophosphate is a key component of regulatory T cell-mediated suppression. Journal of Experimental Medicine, 204: 1303-1310.

9. Brunkow, M.E., Jeffery, E.W., Hjerrild, K.A., Paeper, B., Clark, L.B., Yasayak, S.A., Wilkinson, J.E. Galas D., Ziegler S.F. and Ramsdell F. (2001). Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. Nature Genetics, 27: 68-73.

10. Butcher, S.P. (2003). Target discovery and validation in the post-genomic era. Neurochemical Research, 28: 367-371.

11. Carson,B.D., Lopes, J.E., Soper, D.M. and Ziegler, S.F. (2006). Insights into transcriptional regulation by FOXP3. Frontiers in Bioscience, 11: 1607-1619.

12. Chan S.K., Griffith O.L., Tai I.T., Jones S.J. (2008). Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. Cancer Epidemiology, Biomarkers and Prevention, 17: 543-552.

13. Cline M.S., Smoot M., Cerami E., Kuchinsky A., Landys N., Workman C., Christmas R., Avila-Campilo I., Creech M., Gross B., Hanspers K., Isserlin R., Kelley R., Killcoyne S., Lotia S., Maere S., Morris J., Ono K., Pavlovic V., Pico A.R., Vailaya A., Wang P.L., Adler A., Conklin B.R., Hood L., Kuiper M., Sander C., Schmulevich I., Schwikowski B., Warner G.J., Ideker T., Bader G.D. (2007). Integration of biological networks and gene expression data using Cytoscape. Nature Protocols, 2: 2366-2382.

14. Corthay, A. How do Regulatory T Cells work? (2009). Scandinavian journal of Immunology, 70: 326-336.

15. Dwyer K.M., Gao W., Friedman D., Usheva A., Erat A., Chen J.F., Enjyoji K., Linden J., Oukka M., Kuchroo V.K., Strom T.B., Robson S.C. (2007).

Adenosine generation catalysed by CD39 and CD73 expressed on regulatory T cells mediates immune suppression. Journal of Experimental Medicine, 204: 1527-1265.

16. Di Ianni M., Falzetti F., Carotti A., Terenzi A., Castellino F., Bonifacio E., Del Papa B., Zei T., Ostini R.I., Cecchini D., Aloisi T., Perruccio K., Ruggeri L., Balucani C., Pierini A., Sportoletti P., Aristei C., Falini B., Reisner Y., Velardi A., Aversa F., Martelli M.F. (2011). Tregs prevent GVHD and promote immune reconstitution in HLA-haploidentical transplantation. Blood, 117: 3921–3928.

17. Ezzel, C. (2001). Magic Bullets fly again. Scientific American, 285: 34-41.

18. Forsythe P, Bienenstock J. (2010). Immunomodulation by commensal and probiotic bacteria. Immunological Investigations, 39: 429-448.

19. Furst D.E. (2004), Anakinra: review of recombinant human interleukin-I receptor antagonist in the treatment of rheumatoid arthritis. Clinical Therapeutics, 26: 1960-1975.

20. Galperin, M.Y. and Cochrane, G.R. (2011). The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic. Acids Research, 39 (suppl 1): D1-D6.

21. Gershon, R.K. and Kondo, K. (1970). Cell interactions in the induction of tolerance: the role of thymic lymphocytes. Immunology, 18: 723–737.

22. Gondek D.C., Lu L.F., Quezada S.A., Sakaguchi S., Noelle R.J. (2005). Cutting edge: contact-mediated suppression by CD4+CD25+ regulatory cells involves a granzyme B-dependent, perforin-independent mechanism. Journal of Immunology, 174: 1783-1786.

23. Gottlieb A.B., Kang S., Linden K.G., Lebwohl M., Menter A., Abdulghani A.A., Goldfarb M., Chieffo N., Totoritis M.C. (2004). Evaluation of safety and clinical activity of multiple doses of the anti-CD80 monoclonal antibody, galiximab, in patients with moderate to severe plaque psoriasis. Clinical Immunology, 111: 28-37.

24. Griffin B.D., Verweij M.C., Wiertz E.J. (2010). Herpesviruses and immunity: the art of evasion. Veterinary Microbiology, 143: 89-100.

25. Grossman W.J., Verbsky J.W., Barchet W., Colonna M., Atkinson J.P., Ley T.J. (2004) Human T regulatory cells can use the perforin pathway to cause autologous target cell death. Immunity, 21: 589-601.

26. Haig, D.M.K. (1998). Poxvirus interference with the host cytokine response. Veterinary Immunology and Immunopathology, 63: 149–156.

27. Han J.D., Bertin N., Hao T., Goldberg D.S., Berriz G.F., Zhang L.V., Dupuy D., Walhout A.J., Cusick M.E., Roth F.P., Vidal M. (2004).. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 430: 88-93.

28. Harnett W. and Harnett M.M. (2008) Therapeutic immunomodulators from nematode parasites. Expert Reviews in Molecular Medicine, 10: e18.

29. Hori, S., Nomura, S., Sakaguchi, S. (2003). Control of regulatory T cell development by the transcription factor Foxp3. Science, 299: 1057-1061.

30. Hori, S. (2008). Rethinking the molecular definition of regulatory T cells. European Journal of Immunology, 38: 928-930.

31. Hsu D.H., de Waal Malefyt R., Fiorentino D.F., Dang M.N., Vieira P., de Vries J., Spits H., Mosmann T.R., Moore K.W. (1990). Expression of interleukin-10 activity by Epstein-Barr virus protein BCRF1. Science, 250: 830-832.

32. Huter, E.N., Punkosdy, G.A., Glass, D.D., Cheng, L.I., Ward, J.M. (2008). TGF-beta-induced Foxp3+ regulatory T cells rescue scurfy mice. European Journal of Immunology, 38: 1814-1821.

33. Ideker, T., Ozier, O, Schwikowski, B., Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 18: S233-S240.

34. Kann, M.G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. Briefings in Bioinformatics, 8: 333-346.

35. Kuek A., Hazleman B.L., Ostör A.J. (2007). Immune-mediated inflammatory diseases (IMIDs) and biologic therapy: a medical revolution. Postgraduate Medical Journal, 83: 251-60.

36. Kupershmidt I., Su Q.J, Grewal A., Sundaresh S., Halperin I., Flynn J., Shekar M., Wang H., Park J., Cui W., Wall G.D., Wisotzkey R., Alag S., Akhtari S., Ronaghi M. (2010). Ontology-based meta-analysis of global collections of high-throughput public data. PLoS One, 5: e13066.

37. Lavallée-Adam M., Cloutier P., Coulombe B., Blanchette M. (2011). Modeling contaminants in AP-MS/MS experiments. Journal of Proteome Research. 10: 886-895.

38. Maere S., Heymans K., Kuiper M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. Bioinformatics, 21: 3448-3449.

39. Miyara, M. and Sakaguchi, S. (2011). Human Foxp3+CD4+ regulatory T cells: their knowns and unknowns. Immunology and Cell Biology, 89: 346-351.

40. Morgan M.E., van Bilsen J.H., Bakker A.M., Heemskerk B., Schilham M.W., Hartgers F.C., Elferink B.G., van der Zanden L., de Vries R.R., Huizinga T.W., Ottenhoff T.H., and Toes R.E.. (2005) .Expression of FOXP3 mRNA is not confined to $CD4^{+}CD25^{+}$ T regulatory cells in humans. Human Immunology, 66: 13–20.

41. Mount D.W. and Pandey R. (2005). Using bioinformatics and genome analysis for new therapeutic interventions. Molecular Cancer Therapeutics, 4:1636-1643.

42. Nishizuka, Y., and T. Sakakura. (1969). Thymus and reproduction: sex-linked dysgenesia of the gonad after neonatal thymectomy in mice. Science, 166: 753–755.

43. Ohkura, N., Hamaguchi, M., Sakaguchi, S. (2011). FOXP3^{+} regulatory T cells: control of FOXP3 expression by pharmacological agents. Trends in Pharmacological Sciences, 32: 158-166.

44. Pandiyan, P., Zheng, L., Ishihara, S, Reed, J., Lenardo, M.J. (2007). CD4+CD25+Foxp3+ regulatory T cells induce cytokine deprivation-mediated apoptosis of effector CD4+ T cells. Nature Immunology, 8: 1353-1362.

45. Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. Cell, 144: 864-872.

46. Picirillo, C.A and Shevach, E.M. (2001). Cutting edge: control of CD8+ T cell activation by CD4+CD25+ immunoregulatory cells. Journal of Immunology, 167: 1137-1140.

47. Puccetti P. and Grohmann U. (2007). IDO and regulatory T cells: a role for reverse signalling and non-canonical NF-kappaB activation. Nature Reviews Immunology, 7: 817-823.

48. Ramsey S.A., Gold E.S., Aderem, A. (2010). A systems biology approach to understanding atherosclerosis. EMBO Molecular Medicine, 2: 79-89.

49. Read, S., Malmstrom, V., Powrie, F. (2000). Cytotoxic T lymphocyte-associated antigen-4 plays an essential role in the function of CD25+CD4+ regulatory cells that control intestinal inflammation. Journal of Experimental Medicine, 192: 295-302.

50. Ricke, D.O., Wang, S., Cai, R., Cohen, D. (2006). Genomic approaches to drug discovery. Current Opinion in Chemical Biology, 10: 303-308.

51. Roncarlo, M., and Battaglia, M. (2007). Regulatory T-cell immunotherapy for tolerance to self antigens and alloantigens in humans. Nature Reviews Immunology, 7: 585-598.

52. Sakaguchi, S., Sakaguchi, N., Asano, M. Itoh, M., and Toda, M. (1995). Immunologic self-tolerance maintained by activated T cells expressing IL-2

receptor a-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. Journal of Immunology, 155: 1151–1164.

53. Sakaguchi , S, Wing, K. and Miyara, M. (2007). Regulatory T cells – a brief history and perspective. European Journal of Immunology, 37: S116-123.

54. Sakaguchi S., Yamaguchi T., Nomura T., and Ono M. (2008). Regulatory T cells and immune tolerance. Cell, 133:775-787.

55. Sakaguchi , S., Wing, K., Onishi, Y., Prieto-Martin, P. Yamaguchi, T. (2009). Regulatory Tcells: how do they suppress immune responses? International Immunology, 21: 1105-1111.

56. Seebacher, J. and Gavin, A.-C. (2011). SnapShot: Protein-protein interaction networks. Cell 144: 1000.

57. Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13:2498-2504

58. Shevach, E.M. (2001). Certified Professionals: CD4+CD25+ Suppressor T Cells. Journal of Experimental Medicine, 193: f41–f46.

59. Shevach, E.M. (2006). From vanilla to 28 flavors: Multiple varieties of T regulatory cells. Immunity, 25: 195-201.

60. Spencer J.V., Lockridge K.M., Barry P.A., Lin G., Tsang M., Penfold M.E., Schall T.J.. (2002). Potent immunosuppressive activities of cytomegalovirus-encoded interleukin-10. Journal of Virology, 76:1285-1292.

61. Sprinzak E., Sattath S., Margalit H. (2003). How reliable are experimental protein-protein interaction data? Journal of Molecular Biology, 327: 919-923.

62. Tang Q., Adams J.Y., Tooley A.J., Bi M., Fife B.T., Serra P., Santamaria P., Locksley R.M., Krummel M.F., Bluestone J.A.. (2006) Visualizing regulatory T

cell control of autoimmune responses in non-obese diabetic mice. Nature Immunology, 7: 83-92.

63. Tang Q. and Krummel, M.F. (2006). Imaging the function of regulatory T cells in vivo. Current Opinions in Immunology, 18: 496-502.

64. Tang, Q and Bluestone, J.A. (2008). The Foxp3+ cell: a jack of all trades, master of regulation. Nature Immunology, 9: 239-244.

65. Thornton, A.M. and Shevach, E.M. (1998). CD4+CD25+ immunoregulatory T cells suppress polyclonal T cell activation in vitro by inhibiting interleukin 2 production. Journal of Expeimental Medicine, 188: 287-296.

66. Tran, D.Q., Ramsey, H., Shevach, E.M. (2007). Induction of FOXP3 expression in naïve human CD4+ FOXP3T cells by T-cell receptor stimulation is transforming growth factor-beta dependent but does not confer a regulatory phenotype. Blood, 110: 2983-2990.

67. Troyanskaya, O. (2005). Putting microarrays in a context: Integrated analysis of diverse biological data. Briefings in Bioinformatics, 6: 34-43.

68. Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M., Wodak, S.J. (2011). Interaction databases on the same page. Nature Biotechnology, 29: 391-392.

69. Valenzuela J.G. (2004). Exploring tick saliva: from biochemistry to 'sialomes' and functional genomics. Parasitology, 129: Suppl:S83-94.

70. Vidal, M., Cusick, M.W., Barabasi, A.L. (2011). Interactome networks and human disease. Cell, 144: 986-998.

71. Vincenti F., and Luggen M. (2007). T cell costimulation: a rational target in the therapeutic armamentarium for autoimmune diseases and transplantation. Annual Review Medicine, 58: 347-358.

72. Walker M.R., Kasprowicz D.J., Gersuk V.H., Benard A., Van Landeghen M., Buckner J.H. and Zieglerand S.F. (2003). Induction of FoxP3 and acquisition of

T regulatory activity by stimulated human CD4+CD25+ T cells. Journal of Clinical Investigation, 112: 1437–1443.

73. Wirapati P., Sotiriou C., Kunkel S., Farmer P., Pradervand S., Haibe-Kains B., Desmedt C., Ignatiadis M., Sengstag T., Schütz F., Goldstein D.R., Piccart M., Delorenzi M. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research, 10: R65.

74. Xu, L., Kitani, A., Fuss, I., Strober, W. Cutting edge: regulatory T cells induce CD4+CD25+-FoxP3- cells or are self-induced to become Th17 cells in the absence of exogenous TGF-beta. Journal of Immunology, 178: 6725-6729.

75. Yang, Y., Adelstein, S.J., Kassis, A.I. (2009). Target discovery from data mining approaches. Drug Discovery Today, 14: 147-154.

INTERNET RESOURCES

Bioperl homepage:
www.bioperl.org
(Accessed 18th August 2011)

Bioperl howto page for SearchIO:
http://bioperl.org/wiki/HOWTO:SearchIO
(Accessed 4th August 2011)

Cytoscape website:
www.cytoscape.org
(Accessed 17th August 2011)

EBI InterProScan Sequence Search:
http://www.ebi.ac.uk/Tools/pfa/iprscan/
(Accessed 12th July 2011)

Ingenuity Systems, Inc website:
www.ingenuity.com
(Accessed 18th August 2011)

NCBI batch Entrez:
http://www.ncbi.nlm.nih.gov/sites/batchentrez
(Accessed 8th June 2011)

NCBI FTP listing of blast executables:
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
(Accessed 4th August 2011)

NCBI genome database:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=Genome
(Accessed August 15th 2011)

PSICQUIC: PSI common query interface.
http://www.ebi.ac.uk/Tools/webservices/psicquic/view
(Accessed 14th August 2011)

PSISCORE project website - Scoring method overview:
http://code.google.com/p/psiscore/wiki/Scoring_methods_overview
(Accessed 18th August 2011)

PSISCOREweb, a web-based client for the PSI confidence scoring system:
http://psiscore.bioinf.mpi-inf.mpg.de/
(Accessed 18th August 2011)

Program Parameters for blastall. Tao Tao. User Service, NCBI, NLM, NIH
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall.html#3
(Accessed 5th August 2011)

Program Parameters for formatdb and fastacmd- Two BLAST Database Related Tools.
Tao Tao. User Service, NCBI, NLM, NIH
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html
(Accessed 5th August 2011)

Uniprot FTP listing of databases:
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions (Accessed 6th June 2011)

Uniprot mapping tools:
http://www.uniprot.org/help/mapping?namespace=help&object=mapping&format=tab=batch
(Accessed 27th July 2011)

APPENDIX A – Perl script for BLAST

```
#!/usr/bin/perl  -w


use strict;

use Bio::Tools::Run::StandAloneBlast;
use Bio::SearchIO;
use Bio::SeqIO;


######################
#
# INITIALISE VARIABLES
#
######################

my $in_file = "/home/u046594/blast_db/swissprot_human_seqs.fasta";
my $virus_file = "/home/u046594/blast_db/viral_DB.fasta";

# Set blast parameters

my @params_virus = (-p => 'blastp', -d => 'viral_DB.fasta', -o => 'report.bls', -e => '10');
my @params_human = (-p => 'blastp', -d => 'humanSeqs_fasta.fasta', -o => 'report.bls', -e => '10');

# Instantiate blast objects

my $virus_factory = Bio::Tools::Run::StandAloneBlast->new(@params_virus);
my $human_factory = Bio::Tools::Run::StandAloneBlast->new(@params_human);

my $seqio_object = Bio::SeqIO->new(-file => $in_file);

my $blast_report = "";

open(my $out, '>', 'hits.txt') or die "failed to open output for write: $!";       # open output file for saving hits
open(my $out2, '>', 'BLAST_Result_1.txt') or die "failed to open output for write: $!";  # open output file for saving
results
open(my $out3, '>', 'BLAST_hsp_result.txt') or die "failed to open output for write: $!";  # open output file for saving
results hash
```

```
while (my $seq = $seqio_object->next_seq){
       print "Blasting: ", $seq->id, "...\n";
       my $SEQid = $seq->id;
       my @hum_ids= split('\|', $SEQid);
       my $query_id = $hum_ids[2]; ## Extract Query ID
       print { $out2 } $query_id,"\t";
       my %virus_hit_hash = ();
       my %virus_hsp_hash = ();
       $blast_report = $virus_factory->blastall($seq);
       while (my $result = $blast_report->next_result){
              #print ">", $result->query_name() ,"\n";
              while (my $hit = $result->next_hit){
                     $hit->description =~m/\[(.*)\]/;
                     my $virus = $1;
                     my $virus_hit_id = $hit->accession();

                     print { $out2 } $result->query_accession(),"\t"; #,$result->query_description(),"\t";

                     # output details about the hits   $seq->id, "\t",
                     print { $out2 }  $hit-> description. "\t", $hit-> accession(),"\t",$virus,"\t",$hit-
>significance(),"\t",$hit->num_hsps(),"\t";

                     # output details about the hsps
                     my $length_total =  $result->query_length;
                     while ( my $hsp = $hit->next_hsp ) {
                            if( $hsp->num_identical >= 20 ) {
                                   if ( $hsp->percent_identity >= 20 ) {
                                          print { $out2 } $hsp -> percent_identity(),"\t",$hsp -> length(),"\t",$hsp -
>num_identical(),"\t",$hsp ->num_conserved(),"\t",$hsp ->start('hit'),"\t",$hsp ->end('hit'),"\t",$hsp -
>start('query'),"\t",$hsp ->end('query'),"\t",$hsp -> gaps(),"\t", $hsp->length('query'),"\t";

                                          # output % coverage (No gaps in aligned query seq)
                                          my $length_query =  $hsp->length('query');

                                          my $coverage = ($length_query/$length_total)*100;
                                          print { $out2 }  $coverage,"\t";

                                          # Add to virus hit hash & add to virus hsp info hash

                                          if (not exists $virus_hit_hash{$virus}){
                                                 print ">$virus_hit_id\n";
                                                 $virus_hit_hash{$virus} = $virus_hit_id;
```

```
                                                $virus_hsp_hash{$virus} = $virus_hit_id."\t". $hit-> description. "\t".$hit-
>significance()."\t".$hsp -> percent_identity() . "\t" . $hsp -> length() ."\t".$hsp->length('query'). "\t" .$hsp -
>num_identical(). "\t" .$hsp ->num_conserved(). "\t".$coverage. "\t". $result->num_hits()."\t". $hit-> num_hsps()."\t".
$hsp->rank()."\t";
                                                }
                                        }
                                }
                                last;
                        }
                }
                last;
        }
        foreach my $v (keys %virus_hit_hash){

                open TMP, ">", 'tmp.txt' or die $!;
                print TMP $virus_hit_hash{$v}, "\n";
                system("fastacmd -d viral_DB.fasta -i tmp.txt -p T -o full_seqs_hits.fasta");
                my $seqio_virus = Bio::SeqIO->new(-file => 'full_seqs_hits.fasta');  ## Get virus protein sequence
                my $virus_seq = $seqio_virus->next_seq;

                # Perform reciprocal blast

                my $human_blast_report = $human_factory->blastall($virus_seq);
                while (my $result = $human_blast_report->next_result){
                        print "Blast Report\n";
                        while (my $hit = $result->next_hit){
                                print "Hit\n";
                                my $human_hit_id = $hit->accession;
                                if ($human_hit_id eq $query_id){
                                        print { $out } $query_id, "\t", $human_hit_id, "\t", $v, "\n";
                                        print { $out3 } $human_hit_id, "\t", $result->query_accession,"\t", $v, "\t",
$virus_hsp_hash{$v}, "\n";

                                }
                                last;
                        }
                }
        }


}
```

```
close($out);
close($out2);
close($out3);


       exit;
```

APPENDIX B Perl script for parsing PSICQUIC output. This sample script was for parsing IntAct results, there were slight variations in the scripts for output from other databases, depending on the output format of the results

```
#!/usr/bin/perl

use warnings;
use strict;

my $file = 'scored_INTACT.txt';

open(INFILE, $file) or die "Can't open file: $!\n";
open(my $outfile, '>', 'Intact_parsed_EXTRA.txt');
my @lines = <INFILE>;

my @wanted;

      foreach $_ (@lines) {
            my @columns = split('\t', $_);

                        # first 2 columns:
                        my $col1 = $columns[0];
                        my $col2 = $columns[1];

                        # split at start of uniprot id number:
                        my @split_col1 = split ('[\|]', $col1);
                        my @split_col2 = split ('[\|]', $col2);

                        my $uniprot_num1= $split_col1[0];
                        my $uniprot_num2= $split_col2[0];

                        my @split_id1 = split (':', $uniprot_num1);
                        my @split_id2 = split (':', $uniprot_num2);


                        my $prot_id1 = $split_id1[length(@split_id1)];
                        my $prot_id2 = $split_id2[length(@split_id2)];


                        # interaction type column:
                        my $col12 = $columns[11];

                        # split interaction type columns:
                        my @split_col12 = split ('\(', $col12);

                        my $interaction_long = $split_col12[1];
                        my @interaction_short = split ('\)',
$interaction_long);
                        my $final_interaction= $interaction_short[0];


                        # score column:
                        my $col15 = $columns[14];
                        # split score column:
                        my @split_col15 = split ('MIscore:', $col15);

                        my $score_long = $split_col15[1];
                        my @score_short = split ('\(', $score_long);
                        my $final_score= $score_short[0];
```

```
                        # replace columns 1 and 2 with the uniprot id
number defined as prot_id1, prot_id2
                        # (Q,E means escape any special characters in
the columns, such as |)

                         s/\Q$col1\E/$prot_id1/;
                         s/\Q$col2\E/$prot_id2/;
                         s/\Q$col12\E/$final_interaction/;
                         s/\Q$col15\E/$final_score \n/;

                        # delete rows without uniprot IDs in first 2
columns (these are not PPIs, but chemicals and gene promotor IDs)
                        my @fields = split('\t', $_);
                        if ( $fields[0] =~/^[PQO]/and and $fields[1]
=~/^[PQO]/ ) {
                        push @wanted, $_;
                        print {$outfile} $_;
                        }
      }

exit;
```