Systems biology research on:
1) Computational analysis of Toll-like receptor 3 signaling pathway, and 2) Exploiting proteogenomics for improving *Oryzae sativa* genome annotation.

Mohamed Helmy

Graduate School of Media and Governance

KEIO UNIVERSITY

2009

Systems Biology research on: 1) computational analysis of Toll-like receptor 3 signaling pathway, and 2) Exploiting proteogenomics for improving *Oryzae sativa* genome annotation.

Mohamed Helmy Mahmoud Attia Shehata

KEIO UNIVERSITY

Graduate School of Media and Governance

The thesis Submitted for the fulfillment of the requirements of the degree of Master of Science in the Graduate School of Media and Governance

2009

Abstract

The systemic analysis of the molecular level biological information became indispensable with the vast amounts of data generated by the advanced experimental techniques and high precision measurement tools. Such analysis requires novel multi-disciplinary approaches combining computational analysis with wet-bench experimental methods. Among the numerous available computational methods, mathematical modeling, simulation and tools of high-throughput data analysis provide significant means of novel interpretation of the biological information. Thus, science tends to be discovery-driven beside its normal nature as theory-driven. In this thesis, we present tow systems biology research projects in the systems immunology and proteomics fields, providing an application of systems biology in both theory-driven and discovery-driven science. In the systems immunology project, we analyzed the Toll-like receptor 3 signaling pathway, which plays crucial role in the body against the viral attached, by developing mathematical model simulating its signaling cascades in wildtype and under two knockout conditions. Our analysis revealed the existence of missing cellular events prior to the viral dsRNA recognition and novel pathway required for the MAP kinases activation (chapter 2). Further, in the proteomics project, we performed whole proteome analysis for the undifferentiated cultured cells of rice and used the rice protein, gene and transcript databases to identify the rice proteome. The resultant proteome information was used to verify the rice genome annotation. Through this analysis, we were able to find 210 miss-annotated gene models and provide expression evidence for 38 hypothetical (Chapter 3).

Key Words: Innate Immunity, Toll-like receptor 3, Computational Analysis, Proteogenomics, Genome annotation

Table of Content

List of Tables

List of Figures

Acknowledgment

I believe I was very lucky having Assistant Professor Kumar Selvarajoo, Associate Professor Yasushi Ishihama and Professor Masaru Tomita as my supervisors. I would like to express my gratitude for Assistant Professor Kumar Selvarajoo and Associate Professor Yasushi Ishihama for building my scientific character and for inspiring me with the charming life of the scientific research available at IAB. I benefited a lot from working under their kind guidance where they touch me, patiently and sincerely, the essentials of the scientific research as well as the skills required for conducting productive research. Professor Masaru Tomita represents a special type of supervisors and directors with his understanding of the student's mentality, generous and non-stop support for students. I would like also to present special thanks for Associate Professor Masa Tsuchiya for his support and guidance during my systems immunology project.

I was honored by collaborating with Professor Jun-ichiro Inoue and Dr. Jin Gohda from the University of Tokyo, Professor Shizuo Akira from Osaka University, Professor Koichi Matsuo and Dr. Yasunari Takada from the School of Medicine - Keio University, Professor Naoto Shibuya Meiji University and Dr. Hirofumi Nakagami from RIKEN Plant Science Center.

I am thankful to all faculty members and researchers at IAB past and present. I benefited a lot from the classes, lectures and discussions I had with them. I specially thank, Professor Akio Kanai, Professor Mitsuhiro Itaya, Professor Tomoyoshi Soga, Assistant Professor Martin Robert, Assistant Professor Douglas Murray, Assistant Professor Natsumi Saito and Assistant Professor Arun Krishnan. I am also grateful to Dr. Naoyuki Sugiyama and Mr. Yasuyuki Igarashi for helping me performing my experiments and for performing the in-gel extraction and digestion of the rice cells, respectively. I highly appreciate the help provided by the administrative staff at

IAB and SFC. I specially thank, Akiko Shiozawa, Yu Nakamura, Tomoko Uehara, Naoko Kirisawa, Kenichi Iida, Koichi wada, Haruka Kato and Ayumi Mikami.

I am grateful to all my colleagues and friends in IAB for both friendship and helpful research advices. I specially mention Kentaro Hayashi, Koshi Imami, Kalesh Sasidharan, Cornelia Amariei, Dr. Kosuke Fujishima, Dr. Nozomu Yachie, Vincent Piras, Tsuyoshi Akuzawa, Mio Iwasaki, Atsuko Shinhara, Kaori Takane, Midori Hashimoto, Yoshihiro Toya, Yuka Watanabe and Rainer Machne.

This research was conducted using generous funding from the Ministry of Higher Education and Scientific research - Egypt and its representative in Japan, the Culture, Education and Science Bureau in The Embassy of Egypt in Japan also, with support from Yamagata prefecture, Tsuruoka City - Japan.

I would like to thank my dearly beloved people, Helmy and Mariam, my parents, first fans and continuous supporters. Heba, my wife and Omar, my son who left Egypt where the family and friend and joined me at Japan. I cannot express how happy and proud I am to have those great people near to me all my life giving me their support and love.

Finally, I dedicate this work, with all my gratitude and respect, to my nation hopping to contribute to its development using what I had learned.

Mohamed Helmy
June 30th 2009
Tsuruoka City
Japan

Chapter 1.Introduction

1.1. Systems Biology: approaching systems-level understanding

In the last few decades, research in the biological fields had achieved significant advancements in technologies and methods, resulted in the accumulation of huge amounts of new biological information. Out of many biological fields, molecular biology was one of the fields that received greater attention. Molecular biology, the studying of biology in the molecular level, promised with new understanding for the biological phenomena by utilizing set of techniques that help in characterization, isolation and manipulation of the cell's and the organism's components [Alberts *et al*, 2008]. Expression cloning, polymerase chain reaction (PCR), gel electrophoresis, macromolecule blotting and DNA micro array are some of the key techniques helped the molecular biologists achieve deeper understanding of the characters and functions of the cellular individual components. The rapid advancements in the analytical techniques together with the huge amounts of acquired information raised the need of novel way of studies that provides global understanding of the "system" rather than its individual components. Thus, systems biology research appears as a new field of biological research aims to develop a systems-level understanding [Kitano 2001].

The current attempt of achieving systems-level understanding of the biological systems is not the first attempt. It was preceded by many precursors since the middle of the 20th century. However, the limitations of technologies and the unavailability of the molecular level information made all attempts aiming to system-level understanding in the physiological level rather than the molecular level [Kitano 2001]. For instance, by 1933 Cannon proposed that the biological systems could regulate its internal environment and maintain stable conditions, the process that

named “homeostasis” [Cannon 1933]. By the middle of the 20th century, Norbert Wiener founded the biological cybernetics [Wiener 1948] and two decades later, Von Bertalanffy attempt to form a general theory to the system [Bertalanffy 1968]. Therefore, although systems biology is not the first attempt of accruing system-level understanding for biological systems, it is the first time to approach system-level understanding through knowledge obtained from molecular level and using modern techniques [Kitano 2002].

Systems biology has a very broad scope; however, the main areas of research in the systems biology are i) molecular biology research e.g. genomics, proteomics and recently metabolomics, ii) computational studies e.g. bioinformatics, simulation and databases, iii) analysis of systems’ dynamics, and iv) development of high-precision measurement techniques [Kitano 2001]. Thus, multidisciplinary efforts are required to perform the systems biology research. The understanding and characterization of the system’s individual components are not the only aim of the systems biology research, but also the description of the relationships/interactions between these components. Therefore, it is essential to study the system under different conditions or with certain disruptions or stimulations, which consequently leads to gaining the knowledge required to control the system. Finally, all these knowledge can be used in system’s design or reengineering [Kitano 2007].

Toward the system-level understanding of the biological systems, four phases of research should be accomplished. First phase is the identification of system structure, which should be acquired in both physical level (identification of the system’s individual components and topology) and regulatory level (system’s components interactions and kinetics). Second phase is the analysis of the system’s behavior such as studying the dynamics of the system under different conditions or stimulations e.g. understand the response of the system to the external perturbation. Such

investigations require integrating wet-bench experiments, computational modeling and simulation, and development of sufficient measurement technology .The third phase is system control, which comes as a result of the accumulated knowledge from the two preceding phases. Controlling cancer development, transforming malfunctioning pathways to health pathways or accruing stem cell from a specific cell, are examples of system control. Final phase is system design, where the system structure can be modified or reconstructed in order to acquire certain properties such as bacterial genome engineering [Kitano 2001, Itaya *et al,* 2005 and Kitano 2007].

Scientific research in general can be driven in many ways. Dominantly, science is hypothesis-driven where a question is raised and the current available technologies are used to obtain the necessary information to find the answer [Veenstra 2006]. However, with the current enormous datasets generated using high-throughput technologies such as genome sequencing technologies and mass spectrometry, science tend to be also discovery-driven [Allen 2001]. In the discovery-driven science, also known as data-mining discovery, the information is collected first and then novel finding can be formulated form the information analysis [Smalheiser 2002]. Genomics, proteomics and metabolomics are examples of discovery-driven research as long as high-throughput techniques are used. While, computational methods can be employed in both kinds of research. Systems biology research scope is broad enough to include both kinds of research. Moreover, one research project can combine hypothesis- and discovery-driven approaches [Smalheiser 2002, Veenstra 2006 and Kitano 2007].

In this thesis, we present two systems biology research projects performed in systems immunology (chapter 2) and proteomics (chapter 3), providing examples for application of hypothesis-driven approach and discovery-driven approach, respectively.

1.2. Systems immunology: studying the immune system in an integrative perspective

Immunology is the study of the immune systems in the living organisms. The immune system is the set of organs, tissues, cells and molecular mechanisms that defend the body against infections. It consists of two distinct systems, the innate system and the adaptive system [Kindt *et al,* 2006, Lee and Kim 2007]. The innate immune response is responsible for the rapid and nonspecific detection and clearance of the invading microbes. Therefore, it utilizes set of receptors termed pattern recognition receptors (PRR), such as the Toll-like receptors (TLR) family, which recognize certain molecular patterns associated with the invading pathogens called pathogen-associated molecular patterns (PAMPs) [Krishnan 2007, other book]. While, the consequent adaptive immune response is slower but more specific and provides the host with memory, which allows the adaptive response to be faster and more efficient in the future responses [Kindt *et al,* 2006, Lee and Kim 2007]. The immune system also monitors the environment of the host in order to detect any endogenous abnormal self-identity. Thus, the malfunction or dysregulation of the immune system leads to serious heals problems such as immune-deficiency and autoimmune diseases [Medzhhiov and Janeway 2002, Kindt *et al,* 2006, Lee and Kim 2007].

The Toll-like receptors (TLRs) are the best characterized and the most studied family among all PRRs. To date 13 members where identified in this family in mammalians [Takeda and Akira 2005, Lee and Kim 2007 and Krishnan *et al,* 2007]. The TLRs are type I transmembrane protein working as homodimers or heterodimers or with other PRRs to recognize specific PAMPs. The PAMPs vary from cell-wall component such as bacterial lipopolysaccarides (LPS), microbial protein component such as flagellin or nucleic acids such as viral double-stranded RNA (dsRNA) [Lee and Kim 2007]. Some TLRs are located in the cell surface such as TLR1, TLR2,

and TLR4, mainly to recognize the bacterial components. While, some other members are located on the endocytic compartments, e.g. the endosome, mainly to recognize the viral invaders. The TLR3, TLR7, TLR8 and TLR9 are from the second group [Takeda and Akira 2005, Lee and Kim 2007 and Krishnan *et al,* 2007]. Table 1.1 lists the current 13 members of the TLRs family and their legends.

Although, the immune system was deeply studied and its individual components and functions are well characterized using an efficient reductionist approaches, the overall functioning principles and the regulatory mechanizes cannot be obtained through studying its components as individuals. The strong interactions between the system's components, the pathogen and the surrounding conditions require more comprehensive approaches to study the immune system in an integrated perspective, which is systems immunology [Benoist *et al,* 2006]. Systems immunology is a recent branch from system biology, aims to provide system-level understanding for the immune system. All systems biology tools are applicable in systems immunology studies; however, computational modeling and simulations significantly increased our understanding of the immune response [Covert *et al,* 2005, Selvarajoo 2006a, Santos *et al,* 2007, Selvarajoo *et al,* 2008a, and Helmy *et al,* 2009].

Chapter 2 describes systems immunology research project, where the experimental observations were integrated with computational modeling and simulations to provide better understanding of the TLR3 signaling pathway. Using computational models and experimental activation profiles of the transcription factor nuclear factor-kappaB (NF-κB) and the map kinases Jun N-terminal Kinase (JNK) and p38 mitogen-activated protein kinases (p38) in wildtype, TNF Receptor Associated Factor (TRAF)-6 and NFRSF1A-associated via death domain (TRADD)-deficient murine macrophages, we predicted novel signaling features previously unreported.

e 1.1 the Toll-like receptors and their known legends [Takeda and Akira 2005, Lee and Kim 2007, Krishnan *et al,* 2007].

ıe	Ligand		Source	Example
	Natural	**Synthesized**		
ɪ1	Bacterial Lipoprotein	Pam_3CSK_4 [(1)]	Gram- and Gram+ Bacteria	Wide range of bacteria.
ɪ2	Complement TLR1 in recognition of triacylated bacterial Lipoproteins and TLR6 in recognition of diacylated bacterial lipoproteins.			
ɪ3	Viral dsRNA (double-stranded RNA)	Poly I:C	Virus	SARS, Influenza and Hep
ɪ4	LPS (Lipopolysaccharide)	-	Gram-Negative Bacteria	*Escherichia coli* and *Salm*
ɪ5	Flagellin[(2)]	-	Gram-Negative Bacteria (Flagella)	*Helicobacter pylori, S. typ*
ɪ6	Diacyl lipoproteins	MALP-2 [(3)]	Bacterial lipoproteins	*Mycoplasma fermentans*[(4)]
ɪ7	Viral ssRNA (single-stranded RNA)	poly-uridine (polyU)	Virus	HIV-1
ɪ8	Viral ssRNA (single-stranded RNA)	poly-uridine (polyU)	Virus	HIV-1
ɪ9	unmethylated CpG containing DNA	Synthetic CpG ODN [(5)]	Bacterial DNA	Different types of bacteria
ɪ10[(6)]	Unknown			
ɪ11[(7)]	Profilin-Like protine	-	Protozoan parasite *Toxoplasma gondii*	*Toxoplasma gondii*[(8)]
ɪ12[(7)]	Unknown			
ɪ13[(7)]	Unknown			

ɪ analog of the immunologically active N-Terminal of the bacterial lipoprotein.
protein found in the flagella of Gram-Negative Bacteria.
ell-defined diacylated lipoproteins isolated from *Mycoplasma fermentans.*
ɔssible cofactor contributes to the variation in the time from infection with HIV to the development of AIDS symptoms.
synthetic ologodeoxynucleotides (ODN) containing unmethylated deoxycytosine-deoxyguanosine (CpG) motifs.
ɔes not exist in mouse.
protozoan parasite, the causative agent of toxoplasmosis disease.
oes not exist in human.

1.3. Proteogenomics: Utilizing proteomics in genome annotation

The recent advancements in genome sequencing resulted in enormous growth in sequenced genomes [Wright *et al*, 2009]. On May 2009, the genome online database (http://genomesonline.org/) announced the publication of the 1000th completed genome out of 4852 running genome sequencing projects. However, the genome sequence alone is not sufficient to elucidate the biological functions [Ansong *et al*, 2008a]. Thus, the primary task in any newly sequenced genomes is to attach biological meaning to the sequence; this process is known as genome annotation. The genome annotation usually performed before the genome is deposited in a database or published in a research article. The annotation unite, also known as gene model, is the description of an individual gene and its corresponding transcript and protein products [Koonin and Galperin 2003]. Therefore, annotating the sequenced genome can be performed in two levels i) structural level, identifying the gene structure, and ii) functional level, identifying the biological function of the gene product [Koonin and Galperin 2003, Merrihew *et al,* 2008].

To perform genome annotation, genome sequencing projects usually relay on transcription evidence such as expressed sequence tags (EST), and variety of *in silico* tools for gene finding and protein prediction [Wright *et al*, 2009, Castellana *et al*, 2008, Stanke *et al*, 2008]. Both, transcription evidence and *in silico* prediction, have limitations in genome annotation. Although, cDNA and EST can provide evidence for expression of a predicted gene, they still relay on the un-translated mRNA. Thus, they cannot confirm the expression in the protein level [Wright *et al*, 2009]. While, the *in silico* tools for gene finding and protein prediction vary in their algorithms and accuracy [Coghlan *et al,* 2008, Guigó *et al,* 2006]. Thus, the predicted gene models are usually large and sometimes highly redundant. Moreover, the models are suffering from errors in

reading frames and exon definition [Wright *et al*, 2009, Castellana *et al*, 2008]. These limitations indicated the need of another source of information that helps in correction and confirmation of the predicted gene models.

Proteomics study the details of the expressed proteins and their corresponding peptides in a given sample, including function, structure and sequence [Tyers and Mann 2003]. Thus, the proteome level information can be the desired source of information that complements the traditional genome analysis process. Recently, it became well acknowledged that the inclusion of the proteome data in the genome analysis, results in better genome annotation this new trend called proteogenomics [Desiere *et al,* 2005, Castellana *et al,* 2008, Baerenfaller *et al,* 2008, Power *et al*, 2009, Wright *et al,* 2009, Lasonder *et al,* 2002, Merrihew *et al,* 2008]. Proteogenomics can be defined as, the utilization of large-scale proteome data in genome annotation refinement. Liquid chromatography-mass spectrometry (LC-MS/MS)-based proteomic approaches directly measure the peptides resulted from the expressed proteins and therefore, allow the confirmation/correction of the expressed coding regions in the genomic sequence [Koonin and Galperin 2003, Ansong *et al*, 2008a].

Recent reports show the usefulness of proteogenomics in improving the genome annotation. For instance, *Arabidopsis thaliana* is the most studied plant and the best sequenced and annotated plant genome. However, proteogenomics provided significant additions and corrections to its genome annotation [Castellana *et al,* 2008, Baerenfaller *et al,* 2008]. A genome-scale proteomics study added 57 new gene models to *Arabidopsis* annotation, providing expression evidence for all of them. Moreover, the same study presented functional annotation by flagging the proteins that expressed in one organ only as biomarkers [Baerenfaller *et al,* 2008]. In another proteogenomics study, nearly 13 % of the *Arabidopsis* proteome were deemed incorrect or

missing due to incorrect or missing gene models through finding 778 new protein-coding genes and correcting 695 gene models [Castellana *et al,* 2008]. These significant improvements in the best annotated plant genome "*Arabidopsis* genome" indicate the efficacy of the proteogeomics approaches in improving the genome annotation and promise with significant findings in the incomplete annotated plant genomes, such as rice.

The utility of proteogenomeics in achieving significant improvements in genome annotation had been shown in different organisms as well. In human, Desiere *et al*, demonstrated large scale integration between high-throughput proteomics and the human genome [Desiere *et al,* 2005]. While, Power *et al*, found novel splice isoforms in human platelets using proteomics approach to identify the exon skip events [Power *et al*, 2009]. The *Caenorhabditis elegans* (*C. elegans*) genome annotation was identified, corrected and confirmed using shotgun proteomics by identifying 429 unannotated coding sequences (including 33 pseudogenes), 151 errors in gene models and 254 novel gene models [Merrihew *et al,* 2008]. The same approach is applicable also in genomes of fungi and parasites, the *Aspergillus niger* (the black mold fungus) and *Plasmodium falsiparum* (the malaria parasite) genome annotations were improved using proteogenomic approaches as well [Wright *et al,* 2009, Lasonder *et al,* 2002].

Chapter 3 describes proteogenomics research project, where we performed MS-based shotgun analysis of digested peptides from undifferentiated cultured rice cells. The resultant MS/MS data was compared with the protein, gene and transcript databases of the institute of genome research (TIGR) and then, further processed using bioinformatics tools. Our analysis pointed 210 miss-annotated gene models and provided expression evidence for 38 hypothetical proteins. This work reinforce the importance of including the proteome level information in the genome annotation process.

Chapter 2.Computational analysis of Toll-like Receptor 3 signaling in macrophages

2.1. Introduction

Toll-Like Receptors (TLRs) play a major role in innate immunity, the first line of mammalian defense against invading pathogens and crucial for antigen-specific acquired immunity development [Takeda and Akira 2005, Akira and Takeda 2004]. Upon the recognition of pathogen-associated molecular patterns (PAMPs), such as bacterial lipopolysaccharide (LPS) and viral dsRNA, the 13 currently known TLRs trigger predominantly the activation of MAP kinases and several key transcription factors including nuclear factor-κB (NF-κB), activator protein (AP)-1 and interferon regulatory factor (IRF)- 3 and 7. This results in the induction of numerous proinflammatory cytokines and type I interferons [Krishnan *et al,* 2007, Lee and Kim 2007, Boehme and Compton 2004]. The dysregulation of TLR signaling, therefore, has been attributed to the pathogenesis of major pro-inflammatory illnesses such as the autoimmune diseases [Hulejove *et al,* 2007, Bash 2005].

TLR3, one of the 4 known intracellular members of TLR family, recognizes dsRNA and poly (I:C) [[Krishnan *et al,* 2007, Boehme and Compton 2004, Matsumoto *et al,* 2004] triggers an innate response independent of the adaptor protein Myeloid Differentiation factor 88 (MyD88), which is required for all other TLRs [Matsumoto *et al,* 2004, Johnson *et al,* 2008]. The specificity of TLR3 response is possibly due to the occurrence of an alanine residue in a critical region of its cytoplasmic domain unlike the proline residue utilized by MyD88 found in other TLRs [Boehme and Compton 2004]. Thus, TLR3 initiates its response depending only on the adaptor protein TIR domain-containing adapter-including interferon-β (TRIF) [Akira and Takeda

2004, Matsumoto *et al,* 2003]. The recruitment of TRIF mediates the signaling process through the activation of key transcription factors NF-κB, AP-1, IRF-3 and 7 [Krishnan *et al,* 2007, Boehme and Compton 2004, Matsumoto *et al,* 2004]. Although the signaling molecules and their cascades have been broadly investigated, the dynamic outcome of signal transduction between wildtype and genetic mutations still remains poorly understood.

Here, we began the investigation of TLR3 pathway by literature/database curation of the signaling topology in murine macrophages. Next, we analyzed the temporal experimental data of wildtype, TNF Receptor Associated Factor (TRAF)-6 and NFRSF1A-associated via death domain (TRADD)-deficient murine macrophages with poly (I:C) stimulation [Gohda *et al,* 2004, Pobezinskya *et al,* 2008] by developing a computational TLR3 model based on perturbation-response approach. This approach does not require the detailed reaction kinetics of each reaction in the signaling topology (as in bottom-up approaches), but rather, considers the activation of signaling molecules as linear response events [Selvarajoo *et al,* 2009]. Similar modeling approaches have been previously used to infer important biological network features; inferring feedback control of IKK activity in tumour-necrosis factor (TNF) stimulation [Werner *et al,* 2005], uncovering switching behaviour of MAPK signaling between epidermal growth factor (EGF) and neuronal growth factor (NGF) stimuli [Santos *et al,* 2007], detecting connectivities of reaction molecules [Vance *et al,* 2002], predicting missing molecules in TLR4 signaling [Selvarajoo 2006a] and *signaling flux redistribution* (*SFR*) at pathway junctions [Selvarajoo *et al,* 2008a].

Our TLR3 model simulations were compared with temporal experimental data of NF-κB, JNK and p38 in three experimental conditions; wildtype, TRAF6-deficient and TRADD-deficient macrophages. Collectively, the results suggest i) the existence of novel intermediary steps (e.g.

missing cellular processes, proteins or phosphorylation states) between extracellular poly (I:C) stimulation and intracellular TLR3 binding, and ii) the presence of a novel pathway which is essential for JNK and p38 activation.

2.2. Results and Discussion

2.2.1 Determination of the TLR3 signaling topology in macrophages

TLR3 is expressed in several cell types including macrophages, murine embryonic fibroblasts (MEFs) and dendritic cells (DCs), however, it was not found in B Cells, T Cells and NK cells [Matsumoto *et al,* 2004, Cario *et al,* 2000, Muzio *et al,* 2000, Visintin *et al,* 2001, Matsumoto *et al,* 2003]. Moreover, the experimental observations of TLR3 signaling under various genetic knock-outs show controversial roles of certain molecules in different cell types. For instance, Gohda *et al*. showed that TRAF6 is dispensable for TLR3 induced NF-κB in poly (I:C) stimulated macrophages [Gohda *et al*. 2004], while Jiang *et al*. showed the requirement of TRAF6 to NF-κB activation in MEFs [Jiang *et al,* 2004]. Also, for NF-κB and MAP kinases activation, TRADD is not critical in macrophages, whereas, it is important for MEFs [Pobezinskya *et al,* 2008, Ermolaeva *et al,* 2004]. These data indicate that there is no common topology for TLR3 signaling and, therefore, we cannot combine data obtained from different cell types to create unified TLR3 signaling topology. Instead, independent cell type analysis should be performed.

Here, we investigate the signaling topology for murine macrophages only. It is known that TRIF interacts with TLR3 at the TIR domain and TRAF6 binds to the N-terminal of TRIF [Matsumoto *et al,* 2004, Gohda *et al,* 2004, Sato *et al,* 2003]. However, we found, in murine macrophages, the role of TRAF6 in TLR3 signaling is dispensable [Gohda *et al,* 2004]; the temporal

experimental profiles of MAP kinases (JNK and p38) and NF-κB activation to poly (I:C) stimulation in TRAF6 KO were only slightly reduced compared with wildtype levels (Figure 2.1A-C WT and TRAF6 KO).

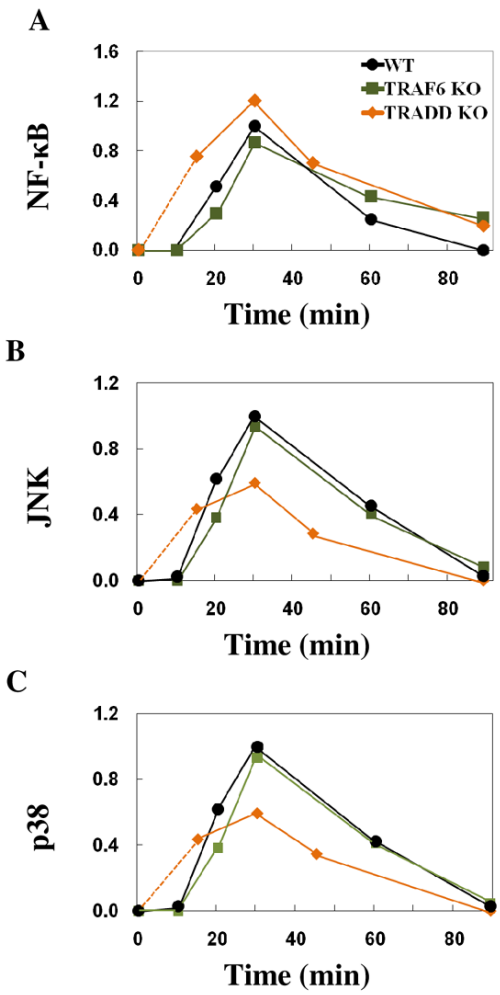


Figure 2.1 NF-κB, JNK and p38 experimental activation profiles. A), B) and C) show experimental activation profiles of NF-κB, JNK and p38, respectively in (WT) (black), TRAF6 KO (green), TRADD KO (orange) obtained from [Gohda *et al,* 2004, Pobezinskya *et al,* 2008]. The activation levels were quantified from the western blots using ImageJ [http://rsbweb.nih.gov/ij/]. The x-axis represents the time in minutes and the y-axis represents the relative activation profile. Note: As TRADD KO data is unavailable at 10 min (earliest at 15 min), we could not observe delayed activation of as noted for WT and TRAF6 KO. Therefore, we used dotted line to connect 0 min and 15 min time points.

Recently, TRADD has been shown to be involved in the TLR3 signaling [Pobezinskya *et al,* 2008, Ermolaeva *et al,* 2004, Chen *et al,* 2003]. TRADD-deficient macrophages showed downregulation of MAP Kinases and a slight upregulation of NF-κB activation (Figure 2.1A-C TRADD KO) [Pobezinskya *et al,* 2008]. RIP1, which binds to TRIF at C-terminal (RHIM domain), interacted with TRADD through death domain (DD) suggesting TRADD's possible involvement in RIP1 ubiquitination [Pobezinskya *et al,* 2008]. Since RIP1 is required for TRIF-dependent signaling in response to poly (I:C) stimulation [Cusson-Hermance *et al,* 2005], TRADD is likely involved in NF-κB activation via RIP1. Putting together, we created a macrophage TLR3 signaling topology (Figure 2.2A).

2.2.2. The prediction of missing intermediate cellular processes in poly (I:C) stimulated macrophages

So far, we have built the signaling topology for TLR3 pathways in macrophages. Next, to understand the complex dynamic interplay of the various intracellular signaling molecules in the regulation of NF-κB, JNK and p38 in poly (I:C) stimulation, we developed a computational model of the TLR3 signaling (see Materials and Methods). Each signaling response (i.e., the rate of activation of each signaling molecule) in the model is represented by $\frac{d\delta \boldsymbol{X}}{dt} = \boldsymbol{J}\delta\boldsymbol{X}$, where $\delta\boldsymbol{X}$ is relative activated concentration of signaling molecules and the parameters (elements of $\boldsymbol{J}$) are chosen to fit the semi-quantitative experimental profiles (e.g, western blots, EMSA, etc.) of NF-κB, JNK and p38 of wildtype macrophages stimulated with poly (I:C) (Figure 2.1A-C WT). As mentioned earlier, TRAF6 KO does not noticeably affect NF-κB, JNK or p38 activation compared to wildtype (Figure 2.1A-C). Hence, in the model, we used a low parameter value

between TRIF and TRAF6 to limit the response flux propagation through TRAF6 such that the removal of TRAF6 in the model will only slightly affect its downstream reactions (see Materials and Methods for details).

We, next, simulated activation of NF-κB, JNK and p38 in wildtype. Our results show the time to reach peak values were 15-20 min earlier than in actual experiments (Figure 2.2B-D). Clearly, this model could not explain the delayed experimental activation of NF-κB, JNK and p38. We can consider time delay processes of signaling events are due to missing molecules/complex formation or spatial movement of molecules [Selvarajoo *et al,* 2009]. The deterministic kinetic evolution equation used in our model, non-linear $\boldsymbol{F}$ in general (Eq.1, Materials and Methods), can include such information by setting total derivative of time to partial derivative in time and space. In other words, Jacobian matrix $\boldsymbol{J}$ can contain temporal and spatial information of the network process. Since we are not performing spatial simulation, time delay response can be lumped as missing molecules/processes in the network. For example, using a TLR4 model, we previously predicted the delayed activation of TRIF-dependent pathways by using a number of additional response reactions representing missing signaling features (molecules/processes, spatial movements or complex formation) in the original network [Selvarajoo 2006a]. Our result was later substantiated by McGettick *et al.,* 2006 [McGettrick *et al,* 2006] who demonstrated two novel signaling molecules (PKCε and TRAM) act upstream of TRIF and, recently, by Kagan *et al.*, 2008 [Kagan *et al,* 2008] who discovered that the internalization of TLR4 into the endosome is required prior to TRIF-dependent pathway activation. Thus, the delayed activation of NF-κB and MAP Kinases in poly (I:C) stimulation could also be due to missing molecules or cellular processes.

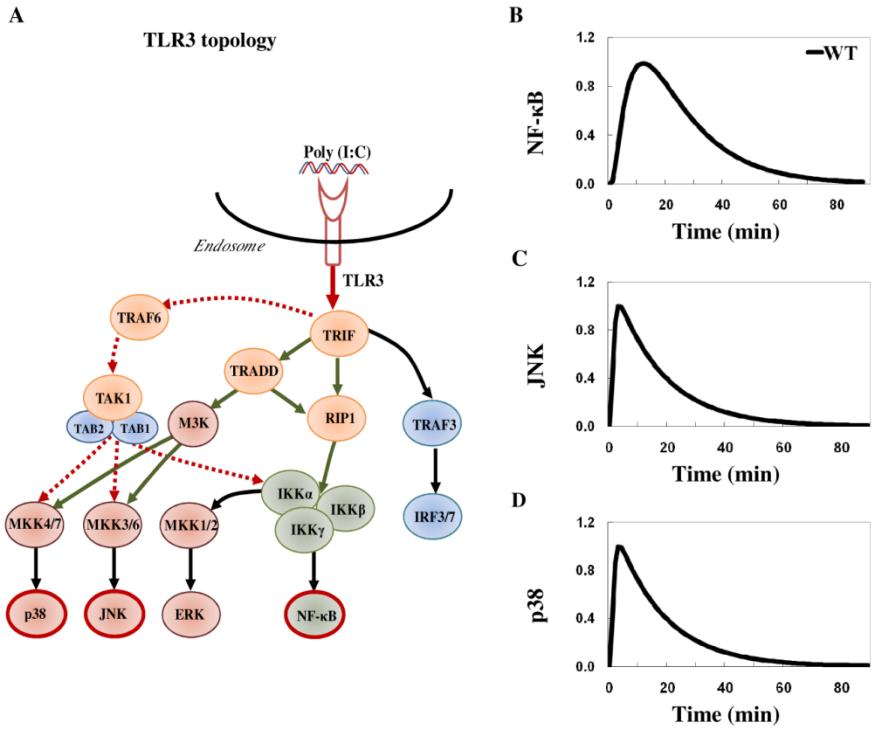


Figure 2.2 Analysis of TLR3 pathway in macrophages. A) Schematic representation of the determined TLR3 pathway topology in macrophages. dsRNA or poly (I:C) stimulated TLR3 triggers TRIF dependant response by the recruitment of TRIF to the cytoplasmic domain of the receptor which then allows RIP1, TRAF6, TBK1 and TRAF3 to bind with TRIF. This results in the activation of MAP kinases (MKK1/2, MKK3/6 and MKK4/7) and IκB kinase complex; MKK1/2, MKK3/6 and MKK4/7 activate ERK, JNK and p38, respectively and IκBα degradation releases NF-κB. TBK1 phosphorylates IRF-3 and 7. ERK, JNK and p38 translocate to the nucleus and activate the transcription factor AP-1, and NF-κB, IRF-3 and IRF-7 translocate to the nucleus. AP-1 and NF-κB bind to the promoter regions of cytokine genes such as *Tnf* and *Il6* while IRF-3, IRF-7 together with NF-κB bind to the promoter region of chemokine genes such as *Cxcl10* and *Ccl5* and induce their transcription. Protein-protein interactions between molecules at the two signaling branches analyzed are highlighted in brown and blue. The dotted lines indicate weak activation (see maintext). B), C) and D) show simulations of NF-κB, JNK and p38 activation, respectively, in wildtype (WT). The *x*-axis represents the time in minutes and the *y*-axis represents the relative activation profile.

To investigate and locate the likely position of the necessary missing intermediary steps in the TLR3 pathway predicted by the model, we further surveyed the literature. TRIF directly bind to the TIR domain of TLR3 and does not require TRAM for its activation [Kagan *et al,* 2008]. Thus, it is unlikely that missing intermediary steps exist between TLR3 and TRIF. Furthermore, TLR3 KO and TRIF KO both showed similar response; abolished activation of NF-κB and MAP Kinases [Yamamoto *et al,* 2003]. These data led us to hypothesize the missing intermediary steps (e.g. signaling molecules or processes) are upstream of TLR3 and we updated our model to begin simulation not from TLR3, but rather from poly (I:C) downwards (Figure 2.3A). We also represented each new uncharacterized molecule/process as a signaling intermediate in our model.

To obtain the delayed activation of NF-κB and MAP kinases in accordance with experimental data of wildtype, that is, null activation till 10 min, peak values around 30 min and reduced activation after 60 min for all molecules (Figure 2.1A-C), we were required to add three intermediary steps (signaling intermediates) (Figure 2.3B-D WT). Having less or more intermediary steps resulted in peak value reaching faster or slower than experimental peak, respectively (data not shown).

While preparing this manuscript, Liu *et al*. demonstrated that TLR3-ectodomains dimerizes before signal propagation [Liu *et al,* 2008]. Our prediction of novel intermediary steps in upstream of TLR3 activation indicates that TLR3-ectodomains dimerization could be one of the missing intermediary steps. The other intermediary steps could indicate the endocytosis of poly (I:C) and its subsequent transport mechanisms to be recognized by TLR3.

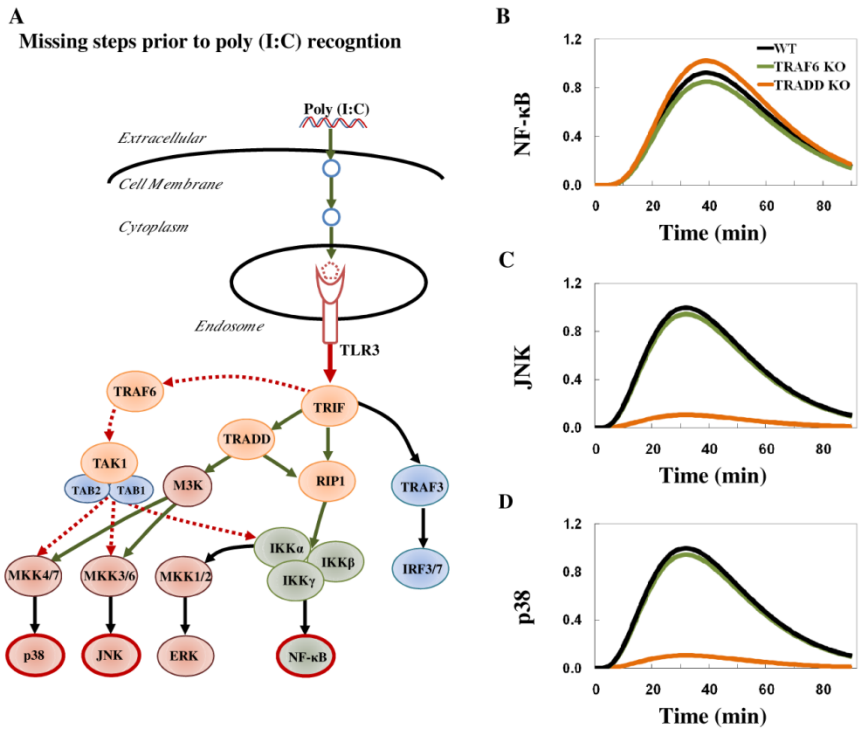


Figure 2.3 Prediction of missing steps prior to poly (I:C)/TLR3 binding. A) Schematic representation of TLR3 model after adding three signaling intermediates upstream of TLR3 representing uncharacterized cellular processes (blue) and TLR3-ectodomain dimerization (red dotted) (see maintext). Note: from our model, it is not possible to equate the three intermediary steps to represent exactly three actual biological events, since spatial transport processes might be one of candidates for the time delay. B), C) and D) show simulations of NF-κB, JNK and p38 activation, respectively, in the wildtype (WT) (black), TRAF6 KO (green), TRADD KO (orange). The *x*-axis represents the time in minutes and the *y*-axis represents the relative activation profile.

2.2.3. The existence of missing pathway for MAP Kinases activation in poly (I:C) stimulation

To further test the predictive capability of our model, we simulated TRAF6 KO and TRADD KO (see Materials and Methods). We wondered whether the same TLR3 model could successfully simulate the activation of NF-κB, JNK and p38 in all three conditions: wildtype, TRAF6 KO and TRADD KO. To create KO condition from wildtype model, we set the reactions of the KO molecule null, while all other model parameters retain their original wildtype values.

The TRAF6 KO simulations matched the experimental data, i.e., the removal of TRAF6 only slightly downregulated NF-κB, JNK and p38 activation (Figure 2.3B-D TRAF6 KO). For TRADD KO, NF-κB activation was slightly increased in experiments (Figure 2.1A). This result was recapitulated by our model which suggests the enhancement of NF-κB activation in TRADD KO is due to *signaling flux redistribution* [Selvarajoo *et al,* 2008a] (Figure 2.3B TRADD KO). However, in contrasts to experimental results, which showed only a small downregulation, simulations of JNK and p38 showed almost abolished activation (Figure 2.3C-D TRADD KO). Thus, our simulation failure suggests, in actual cells, a novel pathway might exist that compensate the loss of MAP Kinases activation in TRADD KO.

Macrophages with TRIF mutation treated with poly (I:C) showed abolishment of MAP kinase ERK activation [Hoebe *et al,* 2009] and all MAP kinases showed similar kinetics in wildtype and TRAF6 KO [Gohda *et al,* 2004]. Thus, we added a pathway from TRIF to MAP kinases into our model (Figure 2.4A and Table 2.1) and adjusted the parameter values between TRIF to RIP1, TRADD and the novel pathway to fit wildtype NF-κB, JNK and p38 activation. We re-

performed the simulations of TRAF6 KO and TRADD KO. The final model simulations matched experimental outcome for all three conditions (Figure 2.4B-D).

To investigate other possible pathways for JNK and p38 activation in poly (I:C) stimulation, reports indicate two other receptors known to recognize dsRNA; the retinoic-acid-inducible protein (RIG)-I and melanoma-differentiation-associated gene (MDA) 5 and, therefore, might be potential candidates [Pindado *et al,* 2007, Kato *et al,* 2007, Takeuchi and Akira 2008]. However, firstly, RIG-I is unable to recognize poly (I:C) [Pindado *et al,* 2007, Takeuchi and Akira 2008] and secondly, MDA5 signaling pathway is unknown to trigger MAP kinases [Takeuchi and Akira 2008]. Furthermore, TRIF mutation showed abolishment of ERK activation [Hoebe *et al,* 2009, Kato *et al,* 2007] and all MAP kinases seem to possess similar kinetics in poly (I:C) stimulation [Gohda *et al,* 2004]. Taken together, these data suggest the predicted pathway for JNK and p38 activation is through TRIF and not by RIG-1 or MDA5.

To find the possible candidates to be involved in the novel pathway, we again surveyed the literature. The TRAF family members, six to-date, are well-known to bind to the TIR domain of TRIF with their C-terminal [Lamkanfi *et al,* 2005, Oganesysn *et al,* 2006, Su *et al,* 2006, Miggin and O'Neill 2006], while two members of the RIP family, RIP1 and RIP3, are found to interact with TRIF through the RHIM domain found in RIP1, RIP3 and TRIF [Meylan *et al,* 2004, Meylan *et al,* 2005]. RIP1, TRAF6 and TRAF3 already exist in the current TLR3 signaling, while TRAF1 was recently found to inhibit the TRIF-mediated signaling [Su *et al,* 2006, Miggin and O'Neill 2006]. Thus, we suggest that RIP3, TRAF2, TRAF4 or TRAF5 may be part of the novel pathway activating JNK and p38.

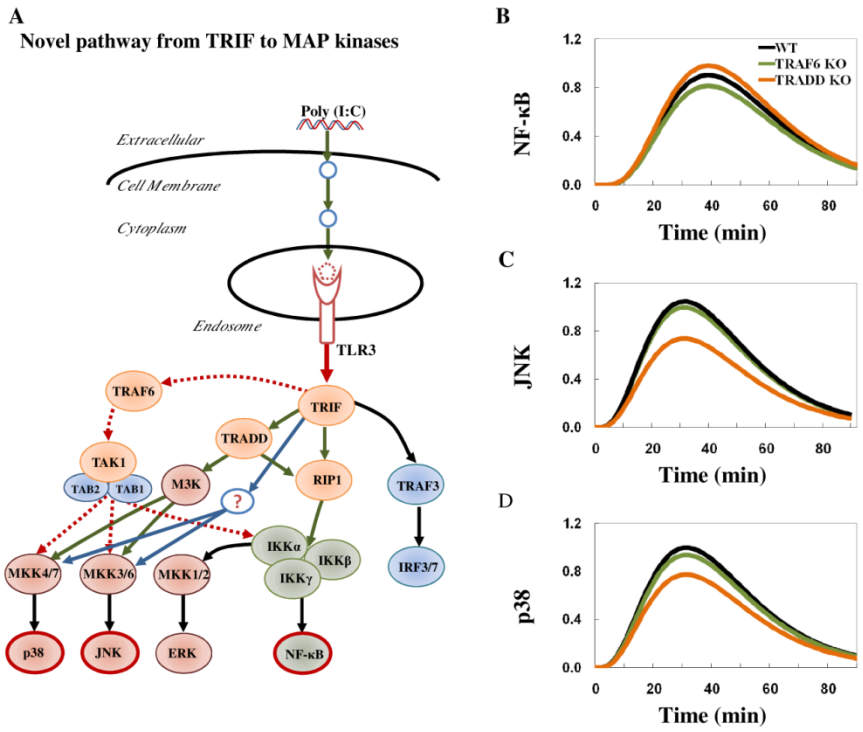


Figure 2.4 A novel pathway is crucial for MAP kinases activation in poly (I:C) stimulated macrophages. A) Schematic representation of the final TLR3 model after adding novel pathway (blue) from TRIF activates MAP kinases. B), C) and D) show simulations of NF-κB, JNK and p38 activation, respectively, of the final macrophages TLR3 model with novel pathway, in wildtype (WT) (black), TRAF6 KO (green) and TRADD KO (orange). The *x*-axis represents the time in minutes and the *y*-axis represents the relative activation profile.

In summary, the analysis of the poly (I:C) stimulation in macrophages reveals the involvement of uncharacterized missing intermediary steps prior to TLR3 activation and the existence of a key pathway from TRIF to JNK and p38, but not NF-κB activation. Although further experimental validation is required to identify the uncharacterized processes and molecules nevertheless, we show that the simple mass-action linear response can represent activation dynamics of the TLR3 signaling, and can be used to predict novel features of complicated immune pathways.

2.3. Materials and Methods

2.3.1. Development of *in silico* TLR3 model using perturbation-response approach

The basic principle behind our perturbation-response approach is to induce a controlled perturbation of a certain (input) reaction molecule of a system, which is kept at steady-state, and monitor the response of the concentration/activation levels of other molecules (output) of the system. By finding a relationship between the input perturbation (e.g. poly (I:C) stimulation) and output response (e.g. JNK, NF-κB activation), the mechanistic properties of the system such as network topology can be uncovered (e.g. novel TRIF-MAP Kinases pathway) without knowing each individual's detailed kinetics.

To illustrate our approach, let us perturb a stable biological network consisting of n molecules from reference (stable) steady-state. The deterministic kinetic evolution equation is

$$\frac{dX_i}{dt} = F_i\left(X_1, X_2, .., X_n\right), i = 1, .., n, \quad [1]$$

where the corresponding vector form of Eq. 1 is $\frac{d\boldsymbol{X}}{dt} = \boldsymbol{F}\left(\boldsymbol{X}\right)$. $\boldsymbol{F}$ is a vector of any non-linear function of the molecules vectors $\boldsymbol{X} = (X_1, X_2, .., X_n)$, which represents activated concentration levels of signaling molecules (for example through phosphorylation, binding concentration of transcription factors to promoter regions etc.). The response to perturbation can be written by $\boldsymbol{X} = \boldsymbol{X}_0 + \delta\boldsymbol{X}$, where $\boldsymbol{X}_0$ is reference steady-state vector and $\delta\boldsymbol{X}$ is relative response from the steady-states ($\delta\boldsymbol{X}_{t=0} = \boldsymbol{0}$).

The use of small perturbation around steady-state leads to important simplification to the evolution equation, which can be highly non-linear (Eq. 1), resulting in the approximation of the

first-order term. In vector form, $\frac{d\delta \boldsymbol{X}}{dt} \cong \left(\left. \frac{\partial F(\boldsymbol{X})}{\partial \boldsymbol{X}} \right|_{X=X_0} \right) \delta \boldsymbol{X}$, noting that zeroth order term $\boldsymbol{F}(\boldsymbol{X}_0)$ = 0 at the steady-state $\boldsymbol{X}_0$ and the *Jacobian* matrix or linear stability matrix, $\boldsymbol{J} = \left. \frac{\partial F(\boldsymbol{X})}{\partial \boldsymbol{X}} \right|_{X=X_0}$. The elements of J are chosen by fitting $\delta \boldsymbol{X}$ with corresponding experimental profiles and knowing the network topology (e.g., activation causality). Hence, the amount of response (flux propagated) along a signaling pathway can be determined using the law of mass flow conservation with *first order mass-action kinetics*, $\frac{d\delta \boldsymbol{X}}{dt} = \boldsymbol{J}\delta \boldsymbol{X}$ (Table 2.1) [Selvarajoo *et al,* 2009, Selvarajoo *et al,* 2008a, Selvarajoo and Tsuchiya 2007]. We considered all signaling response to be linear and sequential up to 90 min after poly (I:C) stimulation (equation [2]) as apparent from the formation-depletion activation profiles (see Experiments, and Results and Discussion). Although there will not be any issue for parameter sensitivity in our system because of linear events, multiple solutions of parameter space, ***J***s, can occur [Selvarajoo *et al,* 2009]. To overcome this, we compare our (wildtype) model with two other (TRAF6 KO and TRADD KO) experimental conditions (see Verification of Model Parameters).

2.3.2. Modeling strategy

We first curated the murine macrophage TLR3 signaling topology from current literature/database sources and by checking with relevant experimental data (Figure 2.5, step 1 and see Results and Discussion) [Gohda *et al,* 2004, Pobezinskya *et al,* 2008, Yamamoto *et al,* 2003, Meylan *et al,* 2005, Kanehisha and Gato 2000]. Next, using the topology, we developed a dynamic model of TLR3 signaling on the E-Cell simulation platform [Takahashi *et al,*2003] to

simulate the dynamics of NF-κB and JNK activation (Figure 2.5, step 2). The parameters of our model (elements of Jacobian matrix ***J***) were estimated by fitting the simulation profiles of NF-κB, JNK and p38 with corresponding activation profiles semi-quantified from immunoblots, obtained from wildtype macrophages [Gohda *et al,* 2004] (Figure 2.5, step 3) (Note that our model works quantitatively if appropriate data is available). If the model successfully fits the wildtype profile of NF-κB and JNK, we accept the model and call it the Reference Model (Figure 2.5, step 4). Otherwise, if the model fails, we adjust the parameters values or model topology until the model is able to simultaneously fit the experimental profiles of NF-κB, JNK and p38 reasonably (Figure 2.5, step 5).

The model simulation results were deemed acceptable by comparing the simulation profile with the experimentally semi-quantitative profile based on three criteria; i) time of activation onset, ii) time to reach peak, and iii) decay profile (Figure 2.5 insert). For example, JNK experimental profile shows null activation until 10 min, peak activation at 30 min and decaying trend until 90 min (Figure 2.1B, WT). In our first simulation of JNK (Figure 2.2C, WT), the time of activation onset was almost instantaneous, peak at around 10 min, and complete decay after 60 min. This simulation, therefore, is considered BAD based on the three criteria which poorly matched (Figure 2.5 insert, upper panel). On the other hand, after modifying the signaling topology by adding novel intermediary steps, the simulation improved on all three criteria and considered GOOD (Figure 2.3C, WT, and Figure 2.5 insert, lower panel).

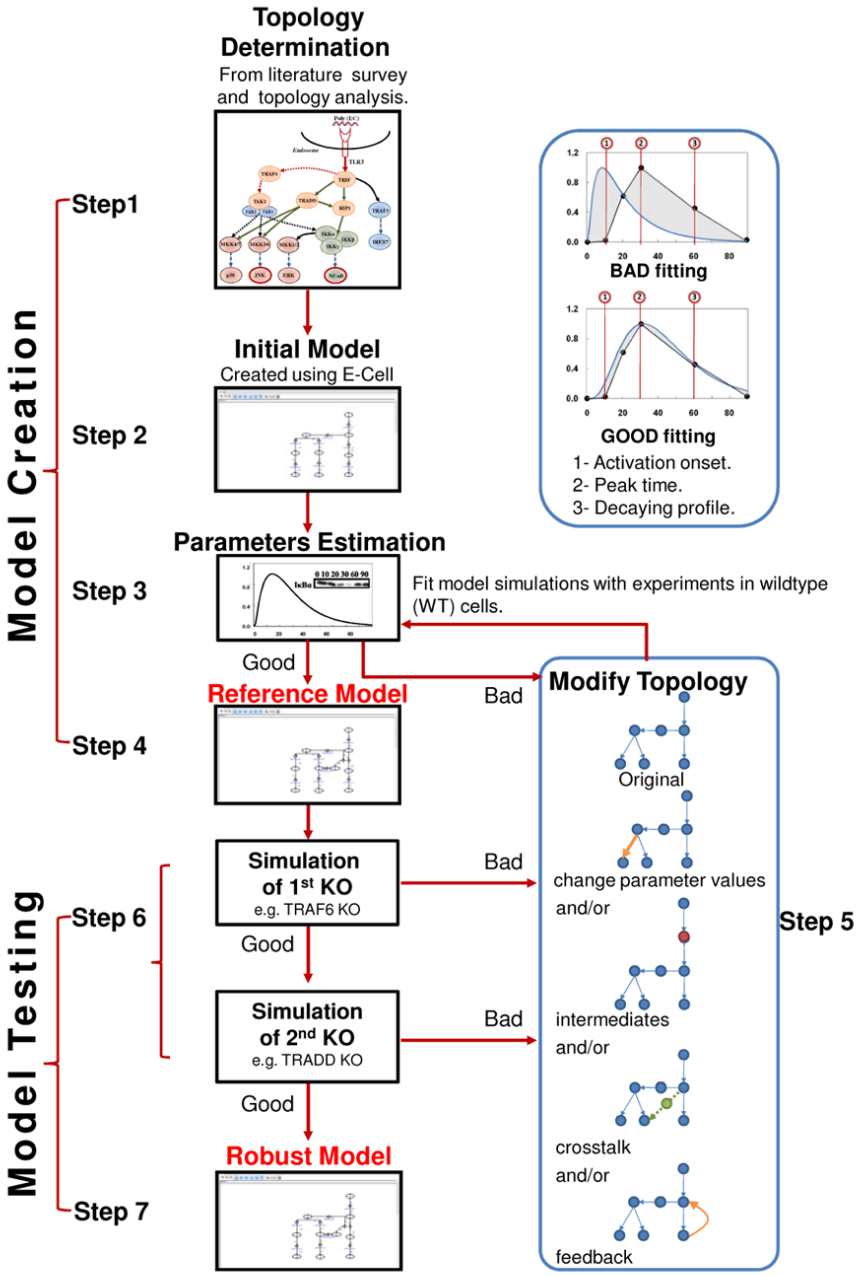


Figure 2.5 The modeling strategy. See maintext for details.

2.3.3. Verification of Model Parameters

There can be several parameter spaces for Jacobian matrix, ***J***s, which could fit wildtype semi-quantitative activation profiles of NF-κB, JNK and p38. To reduce the parameter spaces, we performed an iterative two-mode process of model "creation" and "testing"; "Creation" is to develop and fit our model with published available semi-quantitative profiles of NF-κB and MAP kinases activity in wildtype (Figure 2.5 step 4)."Testing" is to optimize reaction networks and parameter values, obtained from step 4, by simulating the knockout (KO) conditions (TRAF6 KO and TRADD KO) and comparing the results with their respective experimental profiles of NF-κB and MAP kinases activation [Gohda *et al,* 2004, Pobezinskya *et al,* 2008]. Each KO model was generated from the wildtype model by setting the parameters involving the KO molecule null. If our KO simulation does not compare well with experiments, we re-tune the model parameters till the model simulations fit the activation profiles of both wildtype and TRAF6 KO (Figure 2.5 step 6). If unsuccessful, we modify the topology so that a better fit can be obtained, for example, adding novel signaling intermediates to obtain the delayed activation [Selvarajoo 2006a] or crosstalk mechanism to provide an alternative source of activation in the KO condition [Selvarajoo and Tsuchiya 2007, Selvarajoo *et al,* 2008b, Selvarajoo 2006b] (Figure 2.5 step 4, see Results and Discussion). To avoid false prediction, we further check comparing the model simulation against TRADD KO. When the model fails at these steps, it implies that either ***J*** (parameter space) is not correct or the represented topology is insufficient. We repeat the procedure of modifying parameter values and topology until we obtained reasonable predictions for all 3 conditions (wildtype, TRAF6 KO and TRADD KO) at 6 time points (0, 10, 20, 30, 60 and 90 min). At this point, we accepted the model and called it the Robust Model (Figure 2.5 step 7). Therefore, using steps 1-7, we were able to identify missing or

incorrect feature(s) of the TLR3 pathways. The final (robust) TLR3 model's equations and parameter values can be found in Table 2.1.

2.3.4. Final TLR3 Response Model

Our final model consists of 34 molecules with 40 reactions and parameters developed after comparing with dynamics of 3 molecules (NF-κB, JNK and p38) at 6 time points (0, 10, 20, 30, 60 and 90min) for 3 conditions (wildtype, TRAF6 KO and TRADD KO) (Table 2.1).

2.3.5. Experiments on macrophages

We utilized our published time-course experimental data of NF-κB, JNK and p38 of wildtype (model creation) and TRAF6-deficient (model testing) macrophages with 10 μg/ml poly (I:C) stimulation [Gohda *et al,* 2004]. For further model testing, we utilized published experimental profile of NF-κB, JNK and p38 of TRADD-deficient macrophages with similar poly (I:C) stimulation [Pobezinskya *et al,* 2008]. The activation levels of NF-κB, JNK and p38 (Figure 2.1A-C) were quantified from the western blots using ImageJ (http://rsbweb.nih.gov/ij/).

le 2.1 The final *in silico* TLR3 model reactions and parameter values.

Reaction	Formula	Parameter value (1/s)	Remarks
IM1 →IM2	k1*IM1	k1=0.002	
IM2 →IM3	k2*IM2	k2=0.002	Novel intermediates acting upstream of TLR3 representing uncl
IM3 →TLR3	k3*IM3	k3=0.002	
TLR3→TRIF	k4*TLR3	k4= 0.001	TLR3 recruits TRIF.
TRIF→TRAF6	k5* TRIF	k5=0.002	TRAF6 binds to TRIF.
TRIF→ IM4	k6*TRIF	k6=0.004	TRIF activates JNK through Novel pathway
TRIF→RIP1	k7*TRIF	k7=0.007	RIP1 binds to TRIF.
TRIF→TRADD	k8*TRIF	k8=0.002	TRADD binds to TRIF.
TRIF →TRAF3	k9* TRIF	k9=0. 01	TRAF3 binds to TRIF.
TRAF6→ TAB/TAK	k10*TARF6	k10=0.04	TRAF6 binds to TAB/TAK complex.
RIP1→ IKK	k11*RIP1	k11=0.04	RIP1 activated IKK complex.
TRADD→ RIP1	k12*TRADD	k12=0.0001	RIP1 ubiquitination through TRADD.
TRADD → MKKK	k13*TRADD	k13=0.04	TRADD activates JNK and p38 corresponding MKKK.
MKKK → MKK3/6	k14*MKKK	k14=0.04	MKKK activates JNK through activation of MKK3/6.
MKKK → MKK4/7	k15*MKKK	k15=0.04	MKKK activates p38 through activation of MKK4/7.
TAB/TAK →MKK4/7	k16* TAB/TAK	k16=0.03	
TAB/TAK →MKK3/6	k17*TAB/TAK	k17= 0.007	Activation of MAP kinases and IKK via TAB/TAK complex.
TAB/TAK →IKK	k18* TAB/TAK	k18=0.9	
IM4→ MKK3/6	k19* IM4	k19=0.04	The novel pathway activates JNK.
IM4→ MKK4/7	k20* IM4	k20=0.04	The novel pathway activates p38.
IKK → NF-κB/IκB	k21*IKK	k21=0.00167	Phosphorylation of IκBα/NF-κB via IKK complex.
IKK →p105/Tp12	k22*IKK	k22=0.0009	Activation of ERK via IKK.
p105/Tp12→MKK1/2	k23* p105/Tp12	k23=0.003	

NF-κB/IκB →NF-κBc	k24* NF-κB/IκB	k24=0.0333	Release of NF-κB after IκB phosphorylation.
NF-κBc → NF-κBn	k25* NF-κBc	k25=1.0	NF-κB translocate to the nucleus.
NF-κBn → NF-κBdeg	k26* NF-κBn	k26=0.99	NF-κB degradation.
MKK1/2→ERKc	k27*MKK1/2	k27=0.0167	Activation of ERK via MKK1/2.
ERKc→ERKn	k28* ERKc	k28=0.99	ERK translocate to the nucleus.
ERKn→AP-1	k29* ERKn	k29=0.99	ERK activates AP1.
MKK3/6→ JNKc	k30*MKK3/6	k30=0.4	Activation of JNK via MKK3/6.
JNKc→ JNKn	k31*JNKc	k31=1.0	JNK translocate to the nucleus.
JNKn→AP-1	k32* JNKn	k32=0.99	JNK activates AP1.
MKK4/7↔ p38c	k33*MKK4/7	k33=0.4	Activation of p38 via MKK4/7.
p38c → p38n	k34* p38c	k34= 1.0	p38 translocates to the nucleus.
p38n→ AP-1	k35* p38n	k35=0.99	p38 activates AP1.
AP1→AP-1deg	k36* AP1	k36= 0.085	AP1 degradation.
TRAF3→TBK1	k37*TRAF3	k37=0.0001	TRAF3-TBK1 interaction.
TBK1→ IRF-3/7c	k38*TBK1	k38= 0.333	IRF3/7 binds TBK1.
IRF-3/7c→ IRF-3/7n	k39*IRF-3/7c	k39= 0.000167	IRF3/7 translocate to the nucleus.
IRF-3/7n → IRF-	k40*IRF-3/7n	k40=0.0001	IRF3/7 degradation.

Chapter 3.Exploiting proteogenomics for improving *Oryzae sativa* genome annotation

3.1. Introduction

The recent advancements in the experimental techniques resulted in generation of vast amounts of biological data that require special analysis and mining, which therefore, made science more information-driven [Fujishima 2009]. These accumulated amounts of data from different fields, such as genomics and proteomics, require systematic analysis to interpret data from different biological layers in order to reach the system-level understanding of biological systems [Kitano 2007]. To perform the required systematic analysis, new experimental technologies and computational tools were developed, such as annotation pipelines for prokaryotic genomes [Peterson *et al,* 2001, Meyer *et al,* 2003, Van Domselaar *et al,* 2005] and for eukaryotic genomes [Guigo *et al*, 2006, Allen *et al,* 2004, Nielsen et al, 2005], gene-finding tools [Delcher *et al*, 1999, Salzberg *et al*, 1998, Badger *et al,* 1999, Altschul, *et al,* 1997, Altschul, *et al,* 1999, Gish and States 1993], protein-prediction programs [Tatusov *et al*, 1997, Nakai *et al*, 1999, Haft *et al,* 2001, Bateman *et al,* 2004, Hulo *et al,* 2004], and data visualization tools [Stein *et al,* 2002, lewis *et al,* 2002]. However, the data analysis and mining still facing challenges such as data size, differences in data formats and software/algorithms accuracy [Hamady *et al,* 2005, Gupta *et al,* 2008, Ansong *et al,* 2008a].

Among the numerous high throughput experimental methods, the genome sequencing represents a special turning point in the understanding of the biological systems, in general and the genetic blueprints of the organisms, in particular [Ansong *et al,* 2008a, Tanner *et al,* 2007]. Since the year 1995, over 1000 completely sequenced genomes were published and around 4000 other

genome sequencing projects are currently ongoing [GOLD DB http://genomesonline.org/]. In spite of these great efforts in genome sequencing, the biological significance of the sequenced genome cannot be achieved unless the protein-coding genes and their products are accurately identified. Thus, the genome annotation starts to attract the scientists' attention [Koonin and Galperin 2003, Ansong *et al*, 2008a]. The genome annotation is the gene structure and function determination process, and it usually takes place after the genome sequencing and before data deposition to the database or databank [Koonin and Galperin 2003].

In addition to the huge size of data to be analyzed, the annotation process is facing many other challenges [Tanner *et al,* 2007]. In a typical genome annotation work, experimental and computational methods are integrated together to perform the annotation [Koonin and Galperin 2003, Wright *et al*, 2009, Castellana *et al*, 2008, Power *et al,* 2009]. Thus, the genome annotation is highly dependent on the transcription evidence provided by the experiments and the algorithms implemented in the computational tools [Ansong *et al*, 2008a]. However, most of the sequenced genomes are lacking the rich transcriptional evidence, similar to the cDNA library available for the human genome for instance. Even though this information is available it can provide evidence for expression in the transcription level but it cannot confirm the translation into a protein [Ansong *et al*, 2008a, Castellana *et al*, 2008, Baerenfaller *et al,* 2008]. Therefore, the annotation is highly reliant on *de novo* annotations of protein-coding genes performed using gene prediction programs [Koonin and Galperin 2003, Ansong *et al*, 2008a]. In the other hand side, the gene/protein prediction tools had proven their usefulness and utility in the annotation process however, the prediction accuracy vary from tool/algorithm to another and from organism to another according to differences in genome complexity [Ansong *et al,* 2008aCoghlan *et al,* 2008, Guigó *et al,* 2006]. For instance, in the human genome and Arabidopsis genome, the

prediction accuracy was 50% and about 66%, respectively, indicating the need to validate the predicted genes/proteins [Guigó *et al,* 2006, Allen *et al,* 2004].

By the direct measuring of the peptides, LC-MS/MS-based proteomics can provide translation-level expression evidence for the predicted protein-coding genes, in the so-called proteogenomics approach. Proteogenomics is the utilization of large-scale proteome data in genome annotation refinement [Ansong *et al,* 2008a]. This approach seems the best option for identification and confirmation of the protein-coding genes, or at least significant portion of them, in an independent and unambiguous way. This can be achieved through the detection of the naturally occurring proteins (proteomics) and mapping them back to the genome sequence (genomics), in a systematic analysis [Desiere *et al,* 2005, Castellana *et al,* 2008, Baerenfaller *et al,* 2008, Power *et al*, 2009, Wright *et al,* 2009, Lasonder *et al,* 2002, Merrihew *et al,* 2008, Ansong *et al*, 2008a]. In addition to the validation of the predicted gene models in the translation level [Jaff *et al,* 2004, Jaff *et al,* 2005, Wang *et al,* 2005a, Wright *et al,* 2009], proteogenomics can be utilized to reveal other significant genomic features such as finding new gene models [Jaff *et al,* 2004b, Castellana *et al,* 2008, Baerenfaller *et al,* 2008, Merrihew *et al,* 2008], determination of the protein start and termination sites [Tanner *et al,* 2007, Nielsen *et al*, 2005, Mattow *et al,* 2007], finding and verifying splice isoforms in protein level [Tanner *et al,* 2007, Power *et al,* 2009] and verification of hypothetical and conserved hypothetical genes/proteins [Koller *et al,* 2004, Hixson *et al*, 2006, Tanner *et al,* 2007, Ansong *et al*, 2008b]. With the above mentioned efficiency in the improvement of genome annotation, proteogenomics represents a promising approach to be applied in the completed and newly sequenced genomes.

The rice (*Oryza sativa*) is well known as one of the most important crops as almost half of the world population are estimated to be relaying, totally or partially, on it. Moreover, rice

considered one of the model organisms because of its relatively small genome (12 chromosomes and ~370 Mbp) [Sasaki 1999, Wang *et al*, 2005]. Although, the rice whole genome sequence and annotation were published and updated several times (6 builds to date), there are no attempt to include the whole proteome data in genome annotation [Itoh *et al*, 2007, Lin *et al,* 2008, Ouyang *et al,* 2007, Yuan et al, 2005]. Here, we followed systems biology approach, integrating experimental data and bioinformatics tools, to improve the rice genome annotation using whole proteome data. We performed 27 runs of mass spectrometry shotgun analysis (LC-MS/MS) on digested peptides extracted from undifferentiated cultured cells in our lab. The proteins were extracted then digested in-gel and in-solution. The resultant peptide were pre-fractionated using different pre-fractionation methods, strong cation exchange (SCX) [Wu *et al,* 2003] and iso-electric focusing (IEF) [Cargile *et al,* 2004] to broaden our coverage range. Next, Mascot [Perkins *et al,* 1999] search against the TIGR database [Ouyang *et al,* 2007] was used to perform proteins/peptides identification. Through those steps, 5,989 proteins were identified including 69,876 peptides. Then, Mascot was used to search the remaining unidentified spectra against the gene database and transcript database available from TIGR DB. This identification resulted in 577 and 1347 peptides identified from the gene and the transcript databases, respectively. These numbers were reduced through the peptide's score, identification confidence and minimum length (see results and discussion). Thus, we finally got 177 and 697 peptides form the genes and mRNA respectively, indicating existence of miss-annotated gene models. The identified peptides were mapped to the corresponding transcripts and genes resulting in 137 peptides aligned to regions previously considered non-coding regions and 333 intron-spanning peptides. Through this analysis, 210 gene models deemed miss-annotated. In addition, we provided expression evidence for 38 hypothetical and conserved hypothetical proteins 9 of them contain unique

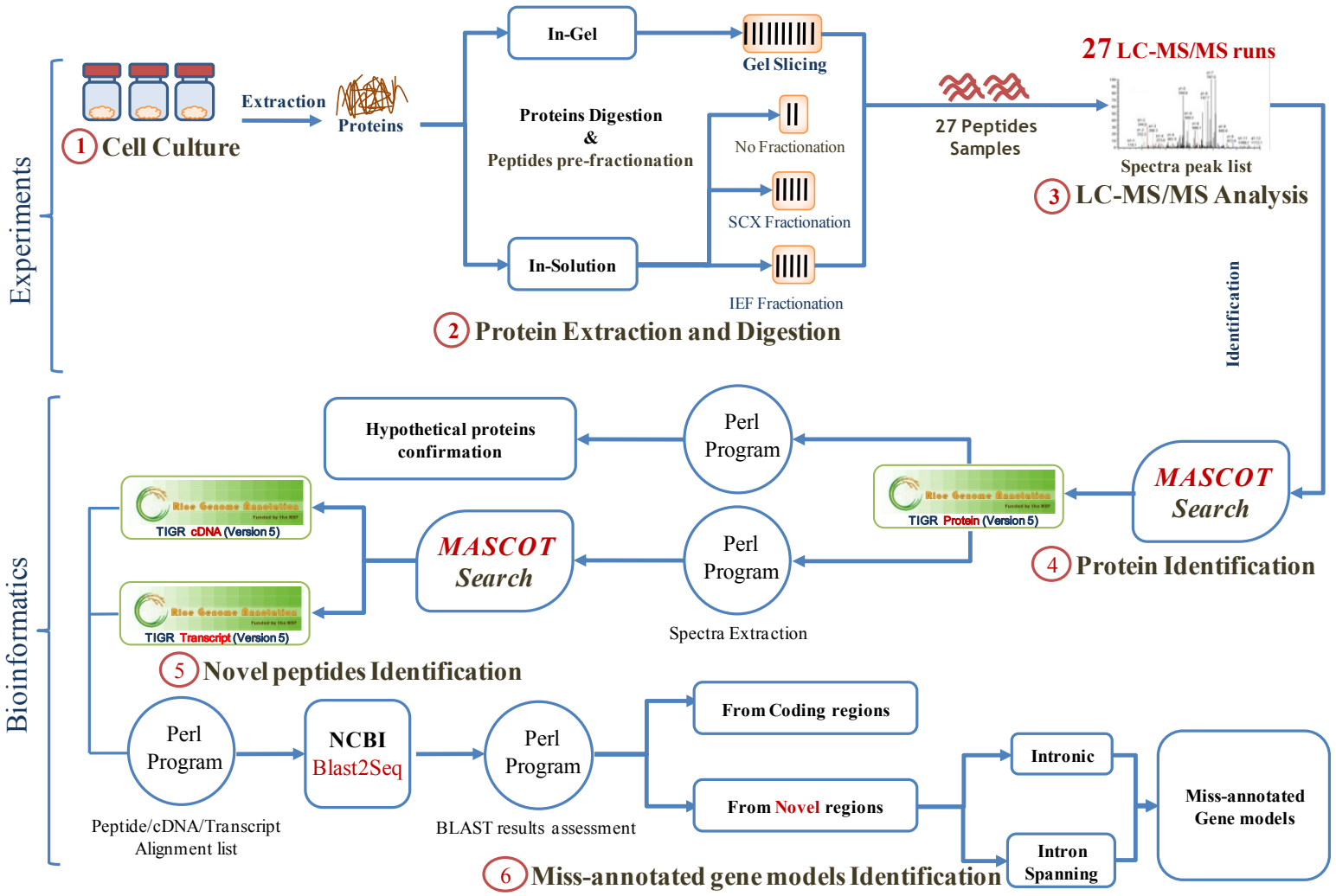
peptides. Our work demonstrates the potential of including proteome level information in genome annotation and supports the inclusion of proteomics module in the genome sequencing projects.

3.2. Results and Discussion

3.2.1. Mass spectrometry-based proteomics of rice cultured cells

The utility of cultured cells in the proteome-wide studies was confirmed, for instance, Baerenfaller *et al* demonstrated the cultured cells proteome covered ~70% of the differentiated organs proteome in *Arabidopsis thaliana,* using extensive sampling from different organs and life stages and performing 1,354 MS runs [Baerenfaller *et al*, 2008]. Thus, we started our study by extracting the proteins from undifferentiated cultured cells and then, the extracted proteins were digested in-gel and in-solution. To achieve broad peptides coverage, the resultant peptides were then pre-fractionated using two pre-fractionation methods, SCX [Wu *et al,* 2003] and IEF [Cargile *et al,* 2004] and then, 27 LC-MS/MS runs were performed. Figure 3.1 shows an overview of the workflow.

The obtained mass spectra were compared against the rice protein database of the institute of genome research (TIGR) [Ouyang *et al,* 2007] using Mascot 2.2 [Perkins *et al,* 1999]. A total of 69,876 peptides were identified with identification confidence 95% contains 27,966 unique peptides sequences. The identified peptides were mapped to 5,989 distinct proteins (Figure 3.2). The contribution of the in-gel extracted samples to the final identification looks significantly greater than the in-solution digested samples (Figure 3.2). However, the in-solution portions in the figures represent the proteins and peptides identified uniquely in the in-solution samples. While, the in-gel portion represents the proteins and peptides identified in the in-gel samples and those were identified in both digestion methods.



The analysis workflow. The extracted proteins from the cultured cell were digested and 27 LC-MS/MS runs were performe ɔtides. The mass spectra were compared against the protein database then the unidentified spectra were extracted and compared a ɔt databases using Mascot. The identified proteins were analyzed to confirm hypothetical proteins, and the novel peptides were u ɛction.

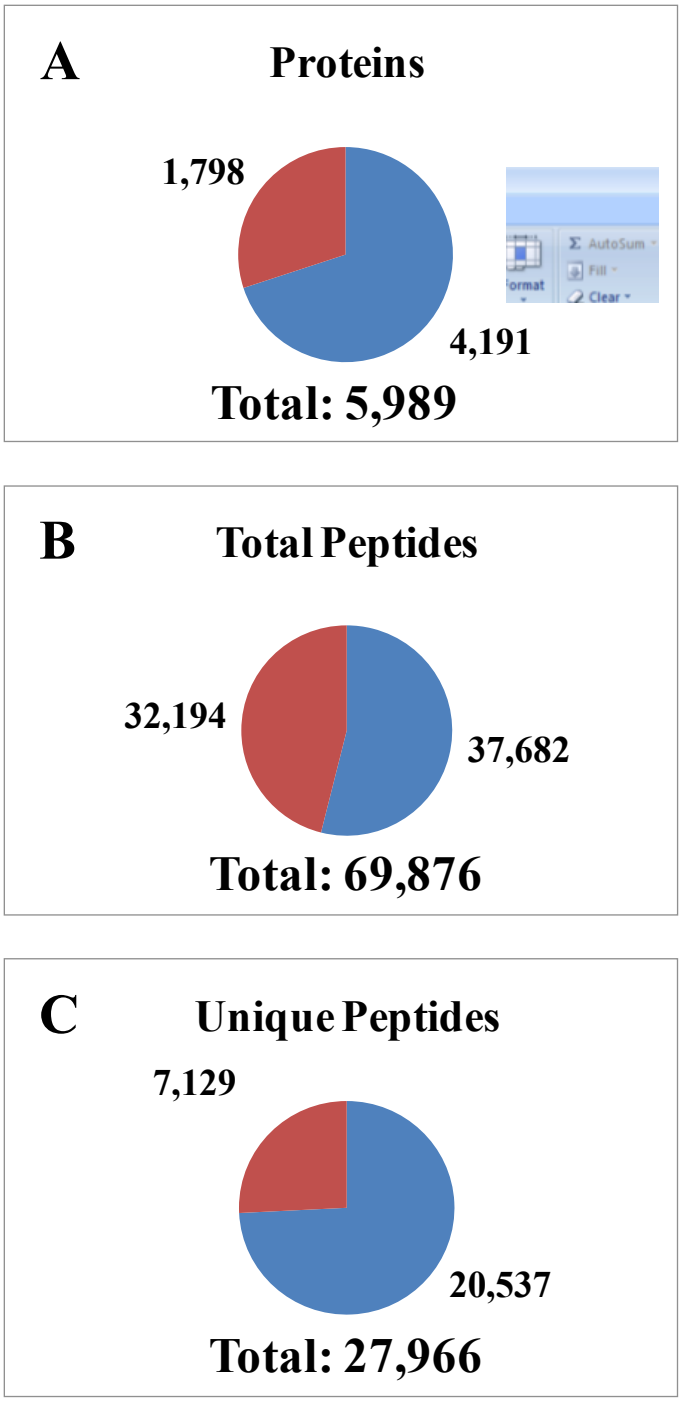


Figure 3.2 Proteome coverage and sample assessment. A), B) and C) represent the identified proteins, total peptides and unique peptides, respectively in the in-solution digestion (red) and in-gel digestion (blue). Note: the in-gel digestion contains the peptides identified in the in-gel digestion and in both digestion methods.

3.2.2. Determination of the shortest accepted peptide's length

The accuracy of the database search and the identification results are crucial for the conclusions of any proteomics study [Hamady *et al,* 2005]. Thus, numerous methods were developed to insure the production of confident identification, accurate database search results and reduced false discovery rate (FDR) such as average peptide scoring (APS) and reverse database searching [Shadforth *et al,* 2005, Tanner *et al,* 2005, Choi and Nesvizhskii 2008]. Although, some peptides identification algorithms, such as Inspect [Tanner *et al*, 2005, Tanner *et al*, 2007], generate peptides starting from three amino acids, the peptide's length still remains one of the false positives sources [Tanner *et al,* 2007]. Peptides with very short length can be mapped to large number of proteins because the short peptide can match similar sequence in a longer peptide. Therefore, the sample with relatively big amount of short peptides will result in higher rate of false positives.

In order to avoid such kind of false positives and increase the accuracy of our identification, we performed an *in silico* study to determine the shortest accepted peptide's length in our analysis. Firstly, we determined the observation range of our machine to know the range of the observable peptides, using the calculated molecular weight and calculated *m/z* in our Mascot identification results. Peptides with molecular weight between 600D and 4000D are within our observation range. Peptides with molecular weight less than 600D or greater than 4000D are unobservable. Next, we created an *in silico* tryptic digested peptides for the entire database and calculated the molecular weight of each peptide. The peptides with molecular weights with our observable range were kept while peptides outside our observation range were disregarded. Samples of 10,000 peptides from each length were created and compared with the whole proteins in the database to calculate the average hits per length. Our aim is the determination of the cutoff point

were the length of the peptide has no effect in the average hits. Our results show the average hits number is decreasing when the peptide's length increases from four amino acids to six amino acids (Figure 3.3.A). Starting from seven amino acids onward, the average hits per length is almost the same (Figure 3.3.A).

On the light of this result, we accept peptides with seven or more amino acids and disregard other peptides. Moreover, we discard proteins identified using short peptides only (less than seven amino acids). Thus, we filtered our identified proteins and peptides based on this rule to minimize the false positives and increase the accuracy of our identification. The filtration resulted in overall average of 4.9% decrees in our identified proteins and peptides (Figure 3.3.B). Table 3.1 list the numbers of identified proteins, peptides and unique peptides before and after filtration and the discarded numbers and their percentages.

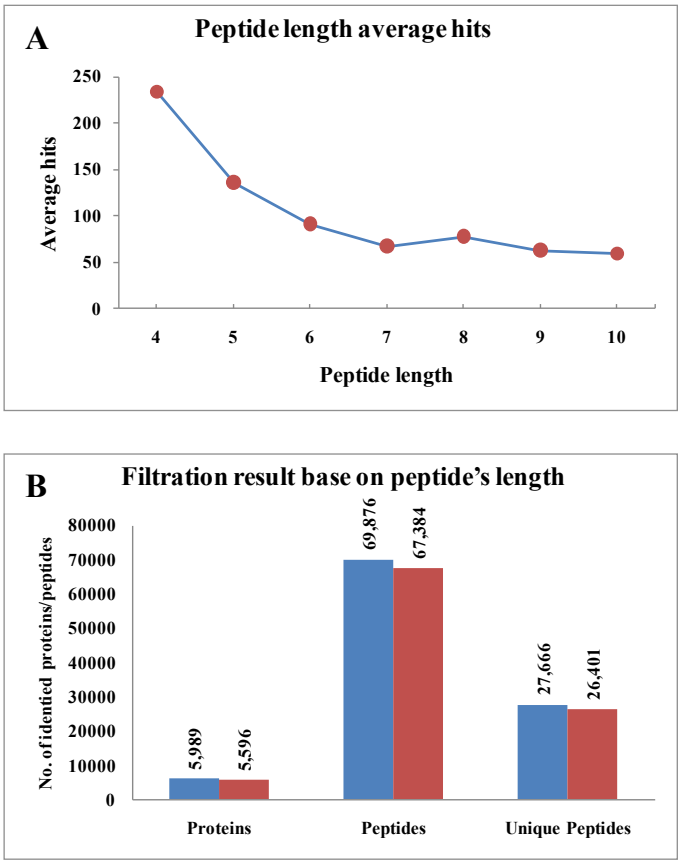


Figure 3.3 Filtration of the identified proteins and peptides based on peptide's length. A) Samples of 10,000 *in silico* tryptic digested peptides were compared against the whole protein database to determine the shortest accepted peptide's length. B) Filtration results show average of 4.9% reduction in the identified proteins and peptides.

Table 3.1 Final numbers of identified proteins and peptides after filtration.

	Total	Filtered	Reduction	Reduction%
Proteins	5,989	5,596	393	6.56%
Total Peptides	69,876	67,384	2,492	3.56%
Unique Peptides	27,666	26,401	1,265	4.57%

3.2.3. Novel peptides identification using gene and transcript databases

So far, our identification results indicate the efficiency of our MS analysis and the sufficient coverage of proteins and peptides. In order to utilize this data in the identification of novel peptides, we used the gene (cDNA) database and transcript (un-spliced mRNA) database available through TIGR rice database [Ouyang *et al,* 2007]. The idea behind using these two databases is that they represent the transcripts and its corresponding spliced model(s). Thus, the identification from both of them should be theoretically similar. Any discrepancy in the peptides identified from them will indicate either new splicing isoform or miss-annotated gene model [Power *et al*, 2009].

Both the gene database and the transcript database are nucleotide sequences database. Searching the nucleotide sequences database is a challenge because of the enormous database size especially with the translated frames and expensive computational processing, which is unaffordable or inapplicable in certain point [Hamady *et al,* 2005, Ansong *et al*, 2008a]. For example, the human proteome database size is about 25 Mb residues while the 6-fram translation of the human genome is about 6Gb residues [Hixson *et al,* 2006]. Moreover, the larger the database sizes the higher false discovery rate, which will require extra analysis to reduce the false discoveries [Ansong *et al,* 2008a]. Thus, the best way to overcome this challenge, with the currently available computational resources and informatics tools, has to be reduction or restriction of the peptides search space [Tanner *et al,* 2007, Ansong *et al*, 2008a]. Therefore, methods were developed to reduce the required search time and computational processing cost by limiting the peptide search space to certain genomic features. For instance, the exon-graph database method, which reduced the human genome database size to 134 Mb and resulted in faster and higher sensitivity search [Tanner *et al,* 2007]. While, the GENQUEST technique

utilizes the isoelectric focusing and accurate mass to reduce the peptide search space [Sevinsky *et al*, 2008].

To reduce our peptide search space while we are searching the gene and transcript database, we followed different approach unlike the currently available. In the developed approaches, the reduction of the search space was achieved by reducing the database size or changing the format of the data by calculating properties that facilitate the search speed and sensitivity [Tanner *et al,* 2007, Sevinsky *et al,* 2008]. While, in our approach we reduced the query size (the MS spectra) which reduced the search time and the processing cost. After searching the protein database, we compared the peptide identification result with the MS spectra data (see materials and methods). All the MS spectra that were used in the peptides identification were excluded and new files, containing the unidentified MS spectra only, were constructed. This resulted in reduction of the MS spectra to be compared with the gene and transcript databases and guarantee the identification of unidentified peptides only. Thus, we do not expect any overlap between peptides identified from protein database and gene/transcript identified ones.

After construction of our new files, that contains the unidentified MS spectra only, we performed benchmarking check to our method. Sample from the new files and their corresponding original files that contain all the MS spectra were compared with the protein database using Mascot. As expected, we did not get any peptide identifications from the new files indicating the complete exclusion of all identified MS spectra. Moreover, the search time required for the new files was relatively shorter (data not shown). Figure 3.4 shows schematic representation of this method.

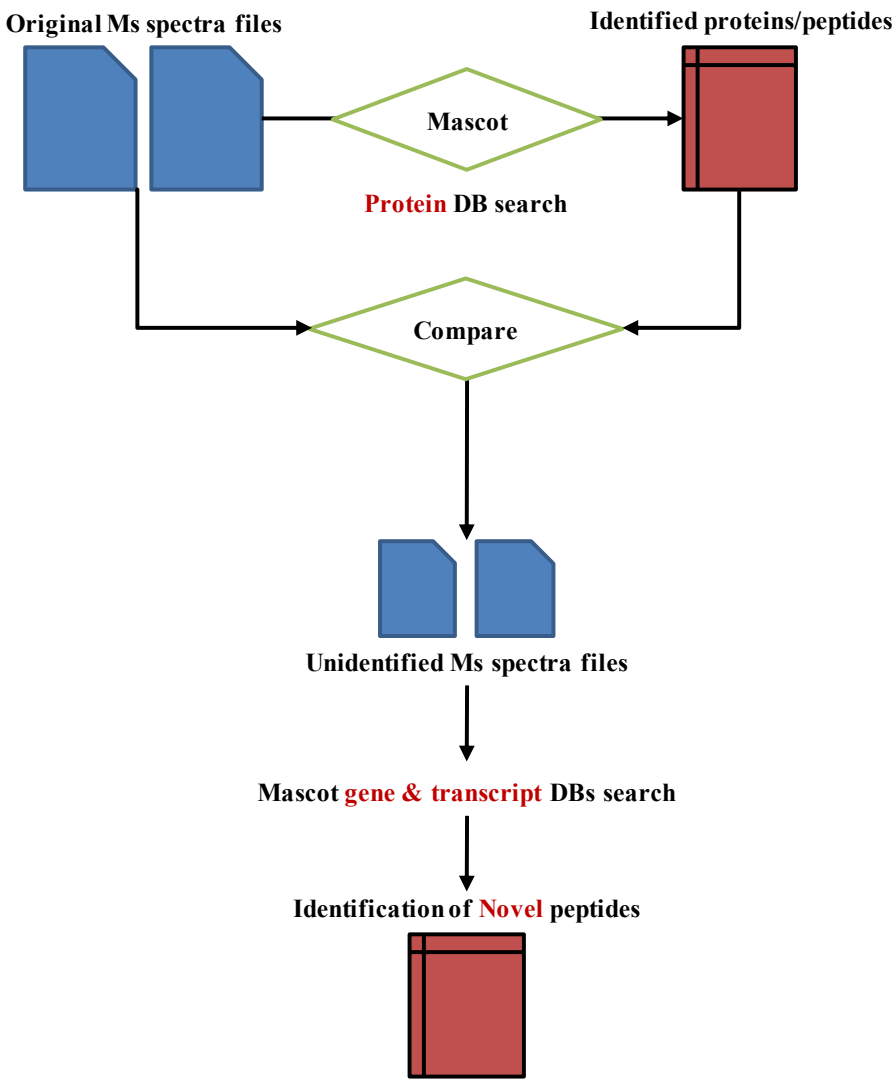


Figure 3.4 Schematic representation of our database search optimization method. The MS spectra were compared against the protein database then the unidentified spectra were extracted and compared with the gene and transcript databases.

The comparison of the new files against the gene and the transcript databases resulted in total of 577 and 1347 peptides identified from the gene and transcript database, respectively. These numbers were reduced to 177 and 697, respectively, through three layers of filtrations based on i) the peptide's score, ii) the peptide's identification confidence (99% or more) and iii) the peptide's minimum length as described above (Figure 3.5).

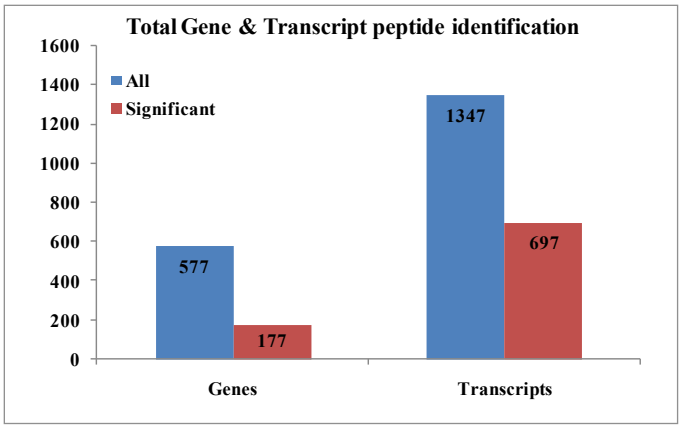


Figure 3.5 Novel peptides identified from gene and transcript databases search. All peptides identified from the gene and transcript databases (blue) were filtered based on peptide's score, identification confidence and minimum length. The remaining deemed significant (red).

3.2.4. Novel peptides alignment to the genes and transcripts reveals new genomic features

The novel peptides identified from the gene and transcript databases were analyzed to find the miss-annotated gene models. Using the gene database identification result, the gene database and the transcript database, we created list of identified peptides, the genes from which they were identified and the transcripts correspond to those genes. Similarly, we created the same list for the transcript database identification result. The two lists were used to perform sequence alignment for the peptides and its corresponding gene and transcript using Basic Local Alignment Search Tool (BLAST) [Altschul *et al,* 1990]. The alignments were performed using local version of NCBI BLAST (blasr2seq) [Tatusova *et al,* 1999] and perl script (Figure 3.6).

The BLAST hit tables of each gene model alignment (peptide against transcript and gene against transcript) were then analyzed using perl program developed for this analysis. The program compares the peptide alignment to the transcript with the gene alignment to the same transcript. Thus, the expected results of this comparison are i) the peptide is from an intronic region, ii) the peptide is intron spanning, or iii) the peptide is from a protein-coding region. Each of the three types indicate novel finding. The intronic peptides possibly indicate new coding region or new exon [Ansong *et al*, 2008, Jaffe *et al*, 2004a, Jaffe *et al*, 2004b]. While, the intron spanning peptides can be used in the exon-intron boundaries correction [Ansong *et al*, 2008]. The peptides from known coding regions can be either from miss-annotated protein or from new splice isoform [Ansong *et al*, 2008, Power et al, 2009]. However, the confirmation of the last type requires the inclusion of the protein in the alignment (Figure 3.6).

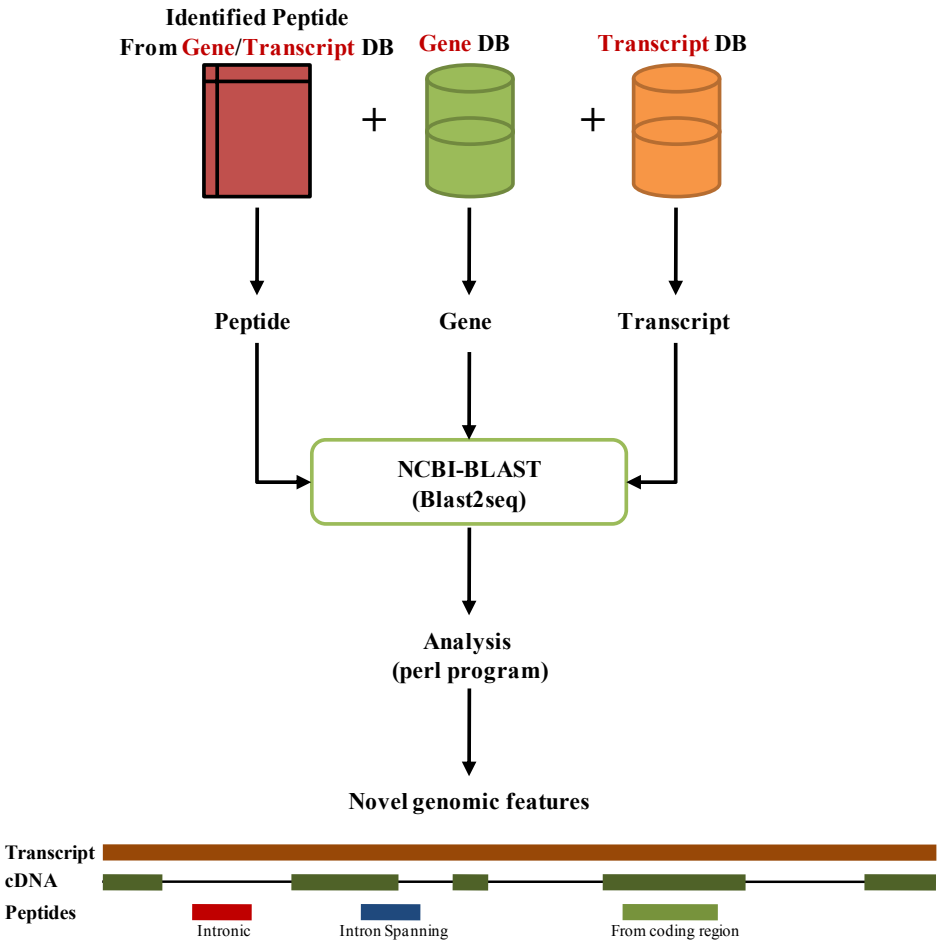


Figure 3.6 The analysis of the novel peptides identified from the gene and transcript databases. A schematic representation of peptides alignments to the corresponding genes and transcripts. The final step explains the new genomic features we could find through this analysis.

The novel peptides identified from the gene and transcript databases had been analyzed independently. The analysis of transcript database identified peptides resulted in 104 peptides from intronic regions, 237 intron spanning peptides and 375 peptides from protein coding regions. Comparing the intronic region and intron spanning peptides clusters, interestingly, 19 peptides were found in both clusters (Figure 3.7.A). This is possibly because of the fact that the identified peptide can be mapped to more than one database entry (protein, gene or transcript) due to the exon sharing between proteins or multiple occurrence of the peptide in the database [Tanner *et al,* 2007]. Thus, the alignment result can differ from one gene model to another for the same peptide. The 322 peptides in the intronic and intron spanning cluster are corresponding to 173 miss-annotated gene models. The miss-annotated gene models also were clustered in three clusters according to the type of miss-annotation i) miss annotated/new coding-region, ii) intron-exon boundary miss-annotation and, iii) models that contains both types of miss-annotations (Figure 3.7.B).

Similarly, the gene database identified peptides had been analyzed revealing 34 peptides from intronic regions, 97 intron spanning peptides, one peptide only in both clusters and 46 peptides from protein coding regions (Figure 3.8.A). Total of 54 gene models were deemed miss-annotated using 131 peptides in the intronic and intron spanning clusters (Figure 3.8.B). Comparing the results of the tow analysis, 451 peptides were used to detect miss-annotated features in the current gene models, new or miss-annotated coding regions (137 peptides) and incorrect intron/exon boundaries (333 peptides) (Figure 3.9.A). While, the comparison between the miss-annotated gene models revealed by the two analysis shows 17 gene model shared between both analyses with total of 210 gene models deemed miss-annotated (Figure 3.9.B). Further analysis is required to re-annotate the miss-annotated gene models.

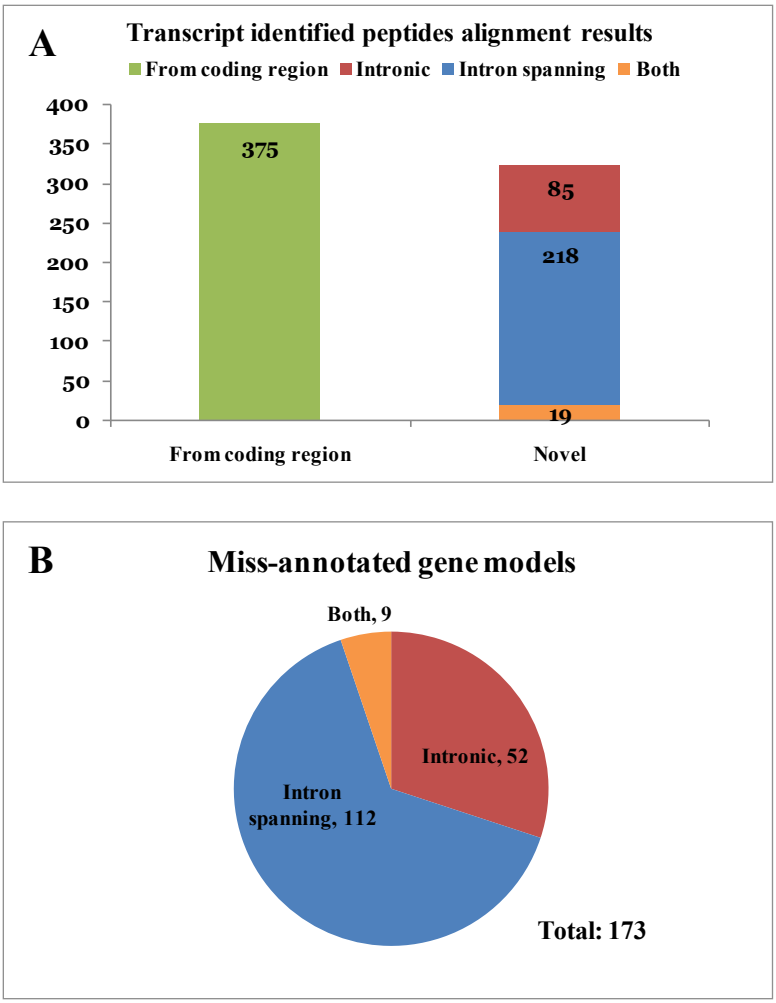


Figure 3.7 Alignment results of peptides identified from the transcript database. A) Shows 322 intronic and intron spanning peptides resulted from the peptides alignments to the corresponding genes and transcripts. B) Shows clustering of the total 173 miss-annotated gene models.

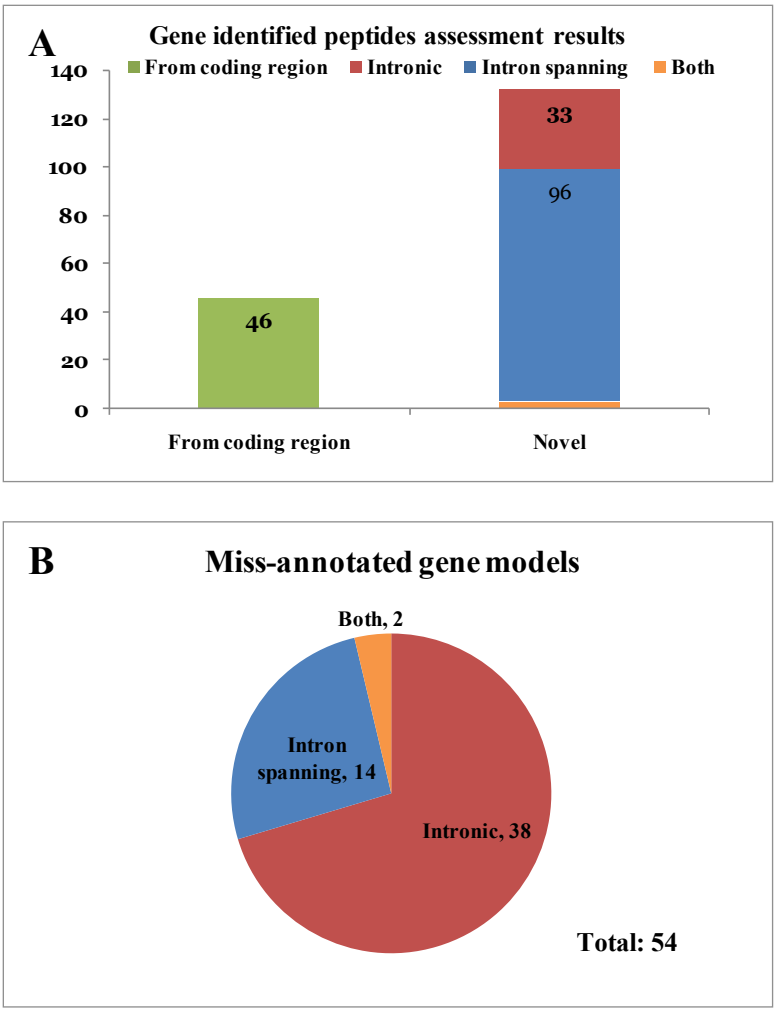


Figure 3.8 Alignment results of peptides identified from the gene database. A) 129 intronic and intron spanning peptides resulted from the peptides alignments to the corresponding genes and transcripts. B) Shows clustering of the total 54 miss-annotated gene models.

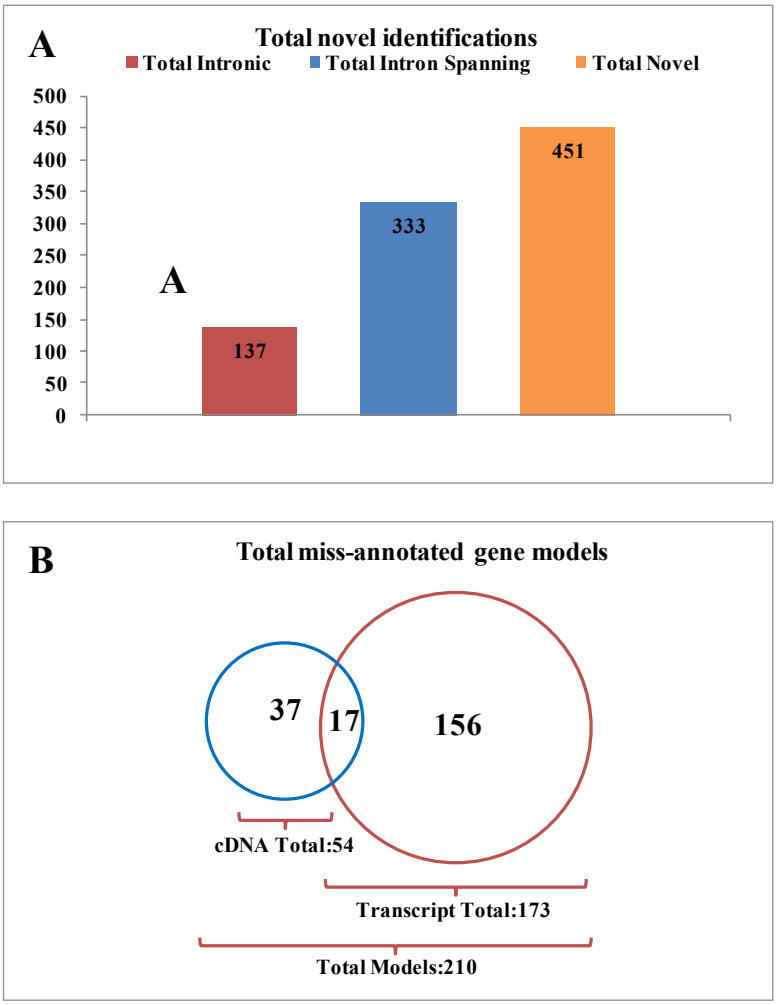


Figure 3.9 Totals of novel peptides and miss-annotated gene models reveled by our analysis.

A) Shows total of novel peptides per cluster. B) Shows total of miss-annotated gene models.

3.2.5. Confirmation of hypothetical proteins

Since the primary genome annotation is a computational step that involves prediction tools and algorithms, the predicted genes/proteins represent around 30~50% of the protein-coding genes in any database. Generally, there are two types of hypothetical proteins, the hypothetical proteins and the conserved hypothetical proteins. The hypothetical proteins are computationally predicted proteins without any known homologs. The conserved hypothetical proteins are, predicted homologous of another hypothetical proteins in a closely related organisms. [Ansong *et al,* 2008a]. Although, the prediction of the protein-coding genes structure is based on the putative start, putative end and upstream promoter region, they still highly erroneous and require experimental validation [Ansong *et al,* 2008a, Tanner *et al,* 2007].

Proteogenomics approach provides an effective way of validating those predicted genes at the translation level through direct measuring the peptides in the LC-MS/MS proteomics [Ansong *et al,* 2008a]. The utility of this approach was demonstrated in many prokaryotic genomes such as *Salmonella typhi* [Ansong *et al*, 2008b], *Salmonellosis typhimurium* [Adkins *et al*, 2006], *Deinococcus radiodurans* [lipton *et al,* 2002] and others [Kolker *et al,* 2004, Kolker *et al,* 2005, Hixson *et al,* 2006, Elias *et al*, 2005]. Later, the same approach was used in higher organisms with complex eukaryotic genomes such as the human genome where 224 hypothetical genes were confirmed using proteogenomics approach [Tanner *et al,* 2007].

The TIGR rice protein database contains 11,950 hypothetical and conserved hypothetical proteins [Ouyang *et al,* 2007]. From our protein database identification result, 333 peptides from 230 proteins annotated as hypothetical proteins were identified, 167 hypothetical proteins and 63 conserved hypothetical proteins. The peptides were filtered based on the same three criteria

mentioned above (see results and discussions), resulting in 38 proteins identified from significant hits.

Since the identified peptide can be assigned to more than one protein due to exon sharing between proteins or multiple occurrence of the peptides within the database [Perkins *et al,* 1999, Tanner *et al,* 2007], we filtered the peptides of the 38 proteins selecting those peptides that were identified in the hypothetical peptides only. We excluded any peptide that matched any other non-hypothetical protein. In addition, we requires more than one peptide with identification confidence more than or equals to 95% or one peptide with identification confidence more than or equals to 99%. The result shows 9 proteins identified from 41 highly significant and uniquely matched peptides. Thus, our work provides the first expression evidence of those proteins in the translation level. Tables 3.2 and 3.3 list the details of the confirmed hypothetical proteins and proteins contains unique peptides, respectively.

Table 3.2 Hypothetical and conserved hypothetical proteins confirmed in this analysis.

Protein	Description
LOC_Os10g31864.1|12010.m65456|	conserved hypothetical protein
LOC_Os01g42990.1|12001.m10562|	conserved hypothetical protein
LOC_Os01g11970.1|12001.m07813|	conserved hypothetical protein
LOC_Os02g40490.1|12002.m09135|	conserved hypothetical protein
LOC_Os01g26832.1|12001.m09121|	conserved hypothetical protein
LOC_Os11g01880.1|12011.m80134|	conserved hypothetical protein
LOC_Os12g38650.1|12012.m07650|	conserved hypothetical protein
LOC_Os05g45910.1|12005.m08716|	conserved hypothetical protein
LOC_Os03g11130.1|12003.m06564|	conserved hypothetical protein
LOC_Os03g60976.1|12003.m10984|	conserved hypothetical protein
LOC_Os11g45410.1|12011.m08373|	conserved hypothetical protein
LOC_Os04g04030.1|12004.m05710|	conserved hypothetical protein
LOC_Os01g08490.1|12001.m07474|	conserved hypothetical protein
LOC_Os12g35910.1|12012.m07384|	conserved hypothetical protein
LOC_Os09g39970.1|12009.m06934|	conserved hypothetical protein
LOC_Os02g21320.1|12002.m07372|	conserved hypothetical protein
LOC_Os01g70280.1|12001.m13076|	hypothetical protein
LOC_Os02g05350.1|12002.m05883|	hypothetical protein
LOC_Os02g18780.1|12002.m07121|	hypothetical protein
LOC_Os01g58300.1|12001.m11974|	hypothetical protein
LOC_Os09g37870.1|12009.m06734|	hypothetical protein
LOC_Os07g15590.1|12007.m06003|	hypothetical protein
LOC_Os06g44630.1|12006.m08990|	hypothetical protein
LOC_Os01g19500.1|12001.m08487|	hypothetical protein
LOC_Os03g24150.1|12003.m07761|	hypothetical protein
LOC_Os03g46900.1|12003.m09705|	hypothetical protein
LOC_Os03g25410.1|12003.m07888|	hypothetical protein
LOC_Os11g03450.1|12011.m04540|	hypothetical protein
LOC_Os11g07410.1|12011.m04937|	hypothetical protein
LOC_Os04g17690.1|12004.m06935|	hypothetical protein
LOC_Os03g25374.1|12003.m07883|	hypothetical protein
LOC_Os01g28850.1|12001.m09325|	hypothetical protein

Continued Table 3.2

LOC_Os11g25600.1|12011.m06481|	hypothetical protein
LOC_Os02g16770.1|12002.m06920|	hypothetical protein
LOC_Os11g06330.1|12011.m04830|	hypothetical protein
LOC_Os07g39500.1|12007.m08192|	hypothetical protein
LOC_Os12g06750.1|12012.m04665|	hypothetical protein

Table 3.3 Hypothetical and conserved hypothetical proteins contain unique peptides.

Protein	Protein description	No. of peptides
LOC_Os01g11970.1|12001.m07813	Conserved hypothetical protein	3
LOC_Os01g26832.1|12001.m09121	Conserved hypothetical protein	2
LOC_Os01g42990.1|12001.m10562	Conserved hypothetical protein	5
LOC_Os02g40490.1|12002.m09135	Conserved hypothetical protein	3
LOC_Os05g45910.1|12005.m08716	Conserved hypothetical protein	3
LOC_Os10g31864.1|12010.m65456	Conserved hypothetical protein	22
LOC_Os11g01880.1|12011.m80134	Conserved hypothetical protein	1
LOC_Os12g38650.1|12012.m07650	Conserved hypothetical protein	1
LOC_Os01g70280.1|12001.m13076	Hypothetical	1

3.3. Materials and Methods

3.3.1. Plant material

The rice (*Oryza sativa* L. cv. Nipponnbare) cell line [Desaki *et al*, 2006] was kindly provided by Prof. Naoto Shibuya (Meiji University), while the cell culture was performed by Dr. Hirofumi Nakagami (RIKEN Plant Science Center) as previously reported [Sugiyama *et al,* 2008].

3.3.2. Protein extraction

3.3.2.1. In-gel extraction

Rice cultured cells (~180 mg) were frozen in liquid nitrogen and then disrupted with a Multi-beads shocker (MB400U; Yasui Kikai). The SDS-PAGE was performed on the disrupted cells with electrophoresis conditions 25mA for 60 min * 20µl in 10 lanes. The gel was later stained with negative gel stain MS kit (Wako, Osaka, Japan). Then, the gel was sliced into 10 slices and each slice sliced again in small cubes ($1mm^3$ cubes). Next, the sample was reduced with dithiothreitol, alkylated with iodoacetamide, and digested with trypsin digestion as described [Ishihama *et al,* 2006]. These digested samples were desalted using StageTips with C18 Empore disk membranes (3 M) [Rappsilber *et al,* 2003]. The whole sample was prepared for the LC-MS/MS analysis.

3.3.2.2 In-solution extraction

Rice cultured cells (~100 mg) were frozen in liquid nitrogen and then disrupted with a Multi-beads shocker (MB400U; Yasui Kikai). The disrupted cells were suspended in 0.1M Tris–HCl (pH 9.0). The sample was transferred to the homogenizer and homogenized. Then, the homogenate was reduced with dithiothreitol, alkylated with iodoacetamide, and digested with Lys-C, followed by dilution and trypsin digestion as described [Saito *et al,* 2006]. These digested samples were desalted using StageTips with C18 Empore disk membranes (3 M) [Rappsilber *et*

al, 2003]. The peptide concentration of the eluates was adjusted to 1.0 mg/ml with 0.1% TFA and 80% acetonitrile.

3.3.3. Peptide pre-fractionation

3.3.3.1. SCX

According to the StageTip fractionation protocol [Rappsilber *et al,* 2007], pre-fractionation using StageTips with a strong cation exchange (SCX) disk was performed for the digested peptides. The SCX-StageTips [Ishihama *et al,* 2006] were used, and 20-500 mM ammonium acetate solutions containing 15% acetonitrile were used for peptides elution, resulting in five fractions. All eluted fractions, including the flow-through fraction, were desalted using means of C18-StageTips.

3.3.3.2. IEF

Another sample from the digested peptides was pre-fractionated using IEF pre-fractionation method by the ZOOM® IEF Fractionator (Invitrogen) according to the manufacturer's protocol. Total of 890 μg of digested peptides were loaded into the ZOOM® IEF Fractionator, 178 μg/chamber, and fractionated at room temperature under standard focusing conditions (100 V for 20 min, 200 V for 80 min, and 600 V for 80 min), resulting in five IEF fractions, enriched in peptides with p*I* values of 3–4.6, 4.6–5.4, 5.4–6.2, 6.2–7, and 7–10.

3.3.4. NanoLC-MS/MS analysis

Total of 27 samples were analyzed using an LTQ-Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany) with a nanoLC interface (Nikkyo Technos, Tokyo, Japan), Dionex Ultimate3000 pump with FLM-3000 flow manager (Germering, Germany), and HTC-PAL autosampler (CTC Analytics, Zwingen, Switzerland). A self-pulled needle (150 mm length, 100 μm i.d., 6 μm opening) packed with ReproSil-Pur C18-AQ materials (3 μm, Dr. Maisch,

Ammerbuch, Germany) was used as an analytical column with "stone-arch" frit. The injection volume was 5 μL, and the flow rate was 500 nL/min. The mobile phases consisted of (A) 0.5% acetic acid and (B) 0.5% acetic acid and 80% acetonitrile. A three-step linear gradient of 5-10% B in 5 min, 10-40% B in 60 min, 40-100% B in 5 min, and 100% B for 10 min was employed throughout this study. A spray voltage of 2400 V was applied. The MS scan range was m/z 300-1500. The top 10 precursor ions were selected in MS scan for subsequent MS/MS scans by ion trapping in the automated gain control (AGC) mode; AGC values of 5.00×10^5 and 1.00×10^4 were set for full MS and MS/MS, respectively. Our system was described previously at Sugiyama *et al,* and Iwazaki *et al* [Sugiyama *et al,* 2008 and Iwazaki *et al,* 2009].

3.3.5. Database search

Peptides and proteins were identified by Mascot v2.2 (Matrix Science, London, U.K.) [Perkins *et al,* 1999] against the TIGR protein database [Ouyang *et al,* 2007], with a precursor mass tolerance of 3 ppm for the in-gel digested samples and 20 ppm for the in-solution ones, and strict specificity allowing for 1 missed cleavage only. Peptides were rejected if the Mascot score was below the 95% confidence limit based on the "identity" score of each peptide, and a minimum of two peptides meeting the criteria was required for protein identification. To increase the identification accuracy and peptide's specificity, we accepted peptides with at least seven amino acids as described above (see results and discussion).

3.3.6. Bioinformatics analysis

The *m*/*z* values of the isotope peaks in the raw data files were converted to the corresponding monoisotopic peaks when the isotope peaks were selected as the precursor ions using and in-house perl script called "mgf creator". The extraction of the unidentified MS spectra was

performed using another perl script called “unidentified spectra extractor (UiSE)” which requires the high performance calculation server at the Institute for Advanced biosciences (Keio University – Japan) to perform the comparison between all MS spectra and the identification results to create new files contains the unidentified MS spectra only. The *in silico* trypsin digestion was performed using perl script provided by Mio Iwazaki (Keio University – Japan). Sequence alignment was performed using local version of NCBI BLAST (blast2seq) windows version [Altschul *et al,* 1990, Tatusova *et al,* 1999] and perl script. We used the default parameters of BLAST except the gapped parameter, we used “gapped = F”. The BLAST hit tables were analyzed using perl scripts.

References

Adkins, J.N., Mottaz, H.M., Norbeck, A.D., Gustin, J.K., Rue, J., Clauss, T.R., Purvine, S.O., Rodland, K.D., Heffron, F. & Smith, R.D., 2006. Analysis of the Salmonella typhimurium proteome through environmental response toward infectious conditions. Mol Cell Proteomics 5(8): 1450-1461.

Akira, S. & Takeda, K., 2004. Toll-like receptor signalling. Nat Rev Immunol 4(7): 499-511.

Allen, J.E., Pertea, M. & Salzberg, S.L., 2004. Computational gene prediction using multiple sources of evidence. Genome Res 14(1): 142-148.

Allen, J.F., 2001. In silico veritas. Data-mining and automated discovery: the truth is in there. EMBO Rep 2(7): 542-544.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215(3): 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17): 3389-3402.

Ansong, C., Purvine, S.O., Adkins, J.N., Lipton, M.S. & Smith, R.D., 2008a. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic 7(1): 50-62.

Ansong, C., Yoon, H., Norbeck, A.D., Gustin, J.K., McDermott, J.E., Mottaz, H.M., Rue, J., Adkins, J.N., Heffron, F. & Smith, R.D., 2008b. Proteomics analysis of the causative agent of typhoid fever. J Proteome Res 7(2): 546-557.

Bach, J.F., 2005. A Toll-like trigger for autoimmune disease. Nat Med 11(2): 120-121.

Badger, J.H. & Olsen, G.J., 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16(4): 512-524.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W. & Baginsky, S., 2008. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 320(5878): 938-941.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. & Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res 32(Database issue): D138-141.

Benoist, C., Germain, R.N. & Mathis, D., 2006. A plaidoyer for 'systems immunology'. Immunol Rev 210: 229-234.

Bertalanffy, L.v., 1968. General system theory. George Braziller, New York, USA

Boehme, K.W. & Compton, T., 2004. Innate sensing of viruses by toll-like receptors. J Virol 78(15): 7867-7873.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts & Walter, P., 2008. Molecular Biology of the Cell: 5th Edition. Garland Scince.

Cannon, B., 1933. Wisdom Of The Body American Public Health Association, USA.

Cargile, B.J., Bundy, J.L., Freeman, T.W. & Stephenson, J.L., Jr., 2004. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J Proteome Res 3(1): 112-119.

Cario, E. & Podolsky, D.K., 2000. Differential alteration in intestinal epithelial cell expression of toll-like receptor 3 (TLR3) and TLR4 in inflammatory bowel disease. Infect Immun

68(12): 7010-7017.

Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V. & Briggs, S.P., 2008. Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A 105(52): 21034-21038.

Chen, N.J., Chio, II, Lin, W.J., Duncan, G., Chau, H., Katz, D., Huang, H.L., Pike, K.A., Hao, Z., Su, Y.W., Yamamoto, K., de Pooter, R.F., Zuniga-Pflucker, J.C., Wakeham, A., Yeh, W.C. & Mak, T.W., 2008. Beyond tumor necrosis factor receptor: TRADD signaling in toll-like receptors. Proc Natl Acad Sci U S A 105(34): 12429-12434.

Choi, H. & Nesvizhskii, A.I., 2008. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res 7(1): 47-50.

Coghlan, A., Fiedler, T.J., McKay, S.J., Flicek, P., Harris, T.W., Blasiar, D. & Stein, L.D., 2008. nGASP--the nematode genome annotation assessment project. BMC Bioinformatics 9: 549.

Covert, M.W., Leung, T.H., Gaston, J.E. & Baltimore, D., 2005. Achieving stability of lipopolysaccharide-induced NF-kappaB activation. Science 309(5742): 1854-1857.

Cusson-Hermance, N., Khurana, S., Lee, T.H., Fitzgerald, K.A. & Kelliher, M.A., 2005. Rip1 mediates the Trif-dependent toll-like receptor 3- and 4-induced NF-{kappa}B activation but does not contribute to interferon regulatory factor 3 activation. J Biol Chem 280(44): 36560-36566.

Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L., 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27(23): 4636-4641.

Desaki, Y., Miya, A., Venkatesh, B., Tsuyumu, S., Yamane, H., Kaku, H., Minami, E. & Shibuya, N., 2006. Bacterial lipopolysaccharides induce defense responses associated

with programmed cell death in rice cells. Plant Cell Physiol 47(11): 1530-1540.

Desiere, F., Deutsch, E.W., Nesvizhskii, A.I., Mallick, P., King, N.L., Eng, J.K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M.G., Kennedy, K.A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J.A., Rawlings, D.J., Samelson, L.E., Shiio, Y., Watts, J.D., Wollscheid, B., Wright, M.E., Yan, W., Yang, L., Yi, E.C., Zhang, H. & Aebersold, R., 2005. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 6(1): R9.

Elias, D.A., Monroe, M.E., Marshall, M.J., Romine, M.F., Belieav, A.S., Fredrickson, J.K., Anderson, G.A., Smith, R.D. & Lipton, M.S., 2005. Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics. Proteomics 5(12): 3120-3130.

Ermolaeva, M.A., Michallet, M.C., Papadopoulou, N., Utermohlen, O., Kranidioti, K., Kollias, G., Tschopp, J. & Pasparakis, M., 2008. Function of TRADD in tumor necrosis factor receptor 1 signaling and in TRIF-dependent inflammatory responses. Nat Immunol 9(9): 1037-1046.

Fernandez-Patron, C., Hardy, E., Sosa, A., Seoane, J. & Castellanos, L., 1995. Double staining of coomassie blue-stained polyacrylamide gels by imidazole-sodium dodecyl sulfate-zinc reverse staining: sensitive detection of coomassie blue-undetected proteins. Anal Biochem 224(1): 263-269.

Fujishima, K., 2009. Systems biology approach toward understanding the central dogma in Archaea. Systems Biology Keio University

Gish, W. & States, D.J., 1993. Identification of protein coding regions by database similarity

search. Nat Genet 3(3): 266-272.

Gohda, J., Matsumura, T. & Inoue, J., 2004. Cutting edge: TNFR-associated factor (TRAF) 6 is essential for MyD88-dependent pathway but not toll/IL-1 receptor domain-containing adaptor-inducing IFN-beta (TRIF)-dependent pathway in TLR signaling. J Immunol 173(5): 2913-2917.

Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E. & Reese, M.G., 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol 7 Suppl 1: S2 1-31.

Gupta, N., Benhamida, J., Bhargava, V., Goodman, D., Kain, E., Kerman, I., Nguyen, N., Ollikainen, N., Rodriguez, J., Wang, J., Lipton, M.S., Romine, M., Bafna, V., Smith, R.D. & Pevzner, P.A., 2008. Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res 18(7): 1133-1142.

Gupta, N., Tanner, S., Jaitly, N., Adkins, J.N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R.D. & Pevzner, P.A., 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res 17(9): 1362-1377.

Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T. & White, O., 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29(1): 41-43.

Hamady, M., Cheung, T.H., Resing, K., Cios, K.J. & Knight, R., 2005. Key challenges in proteomics and proteoinformatics. Progress in proteins. IEEE Eng Med Biol Mag 24(3):

34-40.

Helmy, M., Gohda, J., Inoue, J., Tomita, M., Tsuchiya, M. & Selvarajoo, K., 2009. Predicting novel features of toll-like receptor 3 signaling in macrophages. PLoS ONE 4(3): e4661.

Hixson, K.K., Adkins, J.N., Baker, S.E., Moore, R.J., Chromy, B.A., Smith, R.D., McCutchen-Maloney, S.L. & Lipton, M.S., 2006. Biomarker candidate identification in Yersinia pestis using organism-wide semiquantitative proteomics. J Proteome Res 5(11): 3008-3017.

Hoebe, K., Du, X., Georgel, P., Janssen, E., Tabeta, K., Kim, S.O., Goode, J., Lin, P., Mann, N., Mudd, S., Crozat, K., Sovath, S., Han, J. & Beutler, B., 2003. Identification of Lps2 as a key transducer of MyD88-independent TIR signalling. Nature 424(6950): 743-748.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. & Birney, E., 2005. Ensembl 2005. Nucleic Acids Res 33(Database issue): D447-453.

Hulejova, H., Baresova, V., Klezl, Z., Polanska, M., Adam, M. & Senolt, L., 2007. Increased level of cytokines and matrix metalloproteinases in osteoarthritic subchondral bone.

Cytokine 38(3): 151-156.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. & Bairoch, A., 2004. Recent improvements to the PROSITE database. Nucleic Acids Res 32(Database issue): D134-137.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J. & Frishman, D., 2008. Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics 9: 102.

Ishihama, Y., Rappsilber, J. & Mann, M., 2006. Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics. J Proteome Res 5(4): 988-994.

Itaya, M., Tsuge, K., Koizumi, M. & Fujita, K., 2005. Combining two genomes in one cell: stable cloning of the Synechocystis PCC6803 genome in the Bacillus subtilis 168 genome. Proc Natl Acad Sci U S A 102(44): 15971-15976.

Itoh, T., Tanaka, T., Barrero, R.A., Yamasaki, C., Fujii, Y., Hilton, P.B., Antonio, B.A., Aono, H., Apweiler, R., Bruskiewich, R., Bureau, T., Burr, F., Costa de Oliveira, A., Fuks, G., Habara, T., Haberer, G., Han, B., Harada, E., Hiraki, A.T., Hirochika, H., Hoen, D., Hokari, H., Hosokawa, S., Hsing, Y.I., Ikawa, H., Ikeo, K., Imanishi, T., Ito, Y., Jaiswal, P., Kanno, M., Kawahara, Y., Kawamura, T., Kawashima, H., Khurana, J.P., Kikuchi, S., Komatsu, S., Koyanagi, K.O., Kubooka, H., Lieberherr, D., Lin, Y.C., Lonsdale, D., Matsumoto, T., Matsuya, A., McCombie, W.R., Messing, J., Miyao, A., Mulder, N., Nagamura, Y., Nam, J., Namiki, N., Numa, H., Nurimoto, S., O'Donovan, C., Ohyanagi, H., Okido, T., Oota, S., Osato, N., Palmer, L.E., Quetier, F., Raghuvanshi, S., Saichi, N., Sakai, H., Sakai, Y., Sakata, K., Sakurai, T.,

Sato, F., Sato, Y., Schoof, H., Seki, M., Shibata, M., Shimizu, Y., Shinozaki, K., Shinso, Y., Singh, N.K., Smith-White, B., Takeda, J., Tanino, M., Tatusova, T., Thongjuea, S., Todokoro, F., Tsugane, M., Tyagi, A.K., Vanavichit, A., Wang, A., Wing, R.A., Yamaguchi, K., Yamamoto, M., Yamamoto, N., Yu, Y., Zhang, H., Zhao, Q., Higo, K., Burr, B., Gojobori, T. & Sasaki, T., 2007. Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. Genome Res 17(2): 175-183.

Iwasaki, M., Masuda, T., Tomita, M. & Ishihama, Y., 2009. Chemical cleavage-assisted tryptic digestion for membrane proteome analysis. J Proteome Res 8(6): 3169-3175.

Jaffe, J.D., Berg, H.C. & Church, G.M., 2004a. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4(1): 59-77.

Jaffe, J.D., Stange-Thomann, N., Smith, C., DeCaprio, D., Fisher, S., Butler, J., Calvo, S., Elkins, T., FitzGerald, M.G., Hafez, N., Kodira, C.D., Major, J., Wang, S., Wilkinson, J., Nicol, R., Nusbaum, C., Birren, B., Berg, H.C. & Church, G.M., 2004b. The complete genome and proteome of Mycoplasma mobile. Genome Res 14(8): 1447-1461.

Jiang, Z., Mak, T.W., Sen, G. & Li, X., 2004. Toll-like receptor 3-mediated activation of NF-kappaB and IRF3 diverges at Toll-IL-1 receptor domain-containing adapter inducing IFN-beta. Proc Natl Acad Sci U S A 101(10): 3533-3538.

Johnson, A.C., Li, X. & Pearlman, E., 2008. MyD88 functions as a negative regulator of TLR3/TRIF-induced corneal inflammation by inhibiting activation of c-Jun N-terminal kinase. J Biol Chem 283(7): 3988-3996.

Kagan, J.C., Su, T., Horng, T., Chow, A., Akira, S. & Medzhitov, R., 2008. TRAM couples

endocytosis of Toll-like receptor 4 to the induction of interferon-beta. Nat Immunol 9(4): 361-368.

Kalume, D.E., Peri, S., Reddy, R., Zhong, J., Okulate, M., Kumar, N. & Pandey, A., 2005. Genome annotation of Anopheles gambiae using mass spectrometry-derived data. BMC Genomics 6: 128.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1): 27-30.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. & Kent, W.J., 2003. The UCSC Genome Browser Database. Nucleic Acids Res 31(1): 51-54.

Kato, H., Takeuchi, O., Sato, S., Yoneyama, M., Yamamoto, M., Matsui, K., Uematsu, S., Jung, A., Kawai, T., Ishii, K.J., Yamaguchi, O., Otsu, K., Tsujimura, T., Koh, C.S., Reis e Sousa, C., Matsuura, Y., Fujita, T. & Akira, S., 2006. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. Nature 441(7089): 101-105.

Kindt, T.J., Goldsby, R.A., Osborne, B.A. & Kuby, J., 2006. Kuby Immunology 6th Edition W H Freeman & Co, USA.

Kitano, H., 2001. Foundations of Systems Biology. MIT press, USA.

Kitano, H., 2002. Systems biology: a brief overview. Science 295(5560): 1662-1664.

Kitano, H., 2007. Scientific Challenges in Systems Biology. In: Choi, S. (ed.) Introduction to Systems Biology Humana Press, USA.

Kolker, E., Makarova, K.S., Shabalina, S., Picone, A.F., Purvine, S., Holzman, T., Cherny, T., Armbruster, D., Munson, R.S., Jr., Kolesov, G., Frishman, D. & Galperin, M.Y.,

2004. Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. Nucleic Acids Res 32(8): 2353-2361.

Kolker, E., Picone, A.F., Galperin, M.Y., Romine, M.F., Higdon, R., Makarova, K.S., Kolker, N., Anderson, G.A., Qiu, X., Auberry, K.J., Babnigg, G., Beliaev, A.S., Edlefsen, P., Elias, D.A., Gorby, Y.A., Holzman, T., Klappenbach, J.A., Konstantinidis, K.T., Land, M.L., Lipton, M.S., McCue, L.A., Monroe, M., Pasa-Tolic, L., Pinchuk, G., Purvine, S., Serres, M.H., Tsapin, S., Zakrajsek, B.A., Zhu, W., Zhou, J., Larimer, F.W., Lawrence, C.E., Riley, M., Collart, F.R., Yates, J.R., 3rd, Smith, R.D., Giometti, C.S., Nealson, K.H., Fredrickson, J.K. & Tiedje, J.M., 2005. Global profiling of Shewanella oneidensis MR-1: expression of hypothetical genes and improved functional annotations. Proc Natl Acad Sci U S A 102(6): 2099-2104.

Koonin, E. & Galperin, M., 2003. Sequence-Evolution-Function: Computational Approaches in Comparative Genomics. Kluwer Academic Publishers, USA.

Krishnan, J., Selvarajoo, K., Tsuchiya, M., Lee, G. & Choi, S., 2007. Toll-like receptor signal transduction. Exp Mol Med 39(4): 421-438.

Lamkanfi, M., D'Hondt, K., Vande Walle, L., van Gurp, M., Denecker, G., Demeulemeester, J., Kalai, M., Declercq, W., Saelens, X. & Vandenabeele, P., 2005. A novel caspase-2 complex containing TRAF2 and RIP1. J Biol Chem 280(8): 6923-6932.

Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G. & Mann, M., 2002. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 419(6906): 537-542.

Lee, M.S. & Kim, Y.J., 2007. Signaling pathways downstream of pattern-recognition receptors and their cross talk. Annu Rev Biochem 76: 447-480.

Lewis, S.E., Searle, S.M., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M.A., Kaminker, J.S., Matthews, B.B., Prochnik, S.E., Smithy, C.D., Tupy, J.L., Rubin, G.M., Misra, S., Mungall, C.J. & Clamp, M.E., 2002. Apollo: a sequence annotation editor. Genome Biol 3(12): RESEARCH0082.

Lin, H., Ouyang, S., Egan, A., Nobuta, K., Haas, B.J., Zhu, W., Gu, X., Silva, J.C., Meyers, B.C. & Buell, C.R., 2008. Characterization of paralogous protein families in rice. BMC Plant Biol 8: 18.

Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y., Stritmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, K.K., Zhao, R. & Smith, R.D., 2002. Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. Proc Natl Acad Sci U S A 99(17): 11049-11054.

Liu, L., Botos, I., Wang, Y., Leonard, J.N., Shiloach, J., Segal, D.M. & Davies, D.R., 2008. Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320(5874): 379-381.

Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T., 2005. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33(Database issue): D54-58.

Matsumoto, M., Funami, K., Oshiumi, H. & Seya, T., 2004. Toll-like receptor 3: a link between toll-like receptor, interferon and viruses. Microbiol Immunol 48(3): 147-154.

Matsumoto, M., Funami, K., Tanabe, M., Oshiumi, H., Shingai, M., Seto, Y., Yamamoto, A. & Seya, T., 2003. Subcellular localization of Toll-like receptor 3 in human dendritic cells. J Immunol 171(6): 3154-3162.

Mattow, J., Siejak, F., Hagens, K., Schmidt, F., Koehler, C., Treumann, A., Schaible, U.E. & Kaufmann, S.H., 2007. An improved strategy for selective and efficient enrichment of integral plasma membrane proteins of mycobacteria. Proteomics 7(10): 1687-1701.

McGettrick, A.F., Brint, E.K., Palsson-McDermott, E.M., Rowe, D.C., Golenbock, D.T., Gay, N.J., Fitzgerald, K.A. & O'Neill, L.A., 2006. Trif-related adapter molecule is phosphorylated by PKC{epsilon} during Toll-like receptor 4 signaling. Proc Natl Acad Sci U S A 103(24): 9196-9201.

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

Merrihew, G.E., Davis, C., Ewing, B., Williams, G., Kall, L., Frewen, B.E., Noble, W.S., Green, P., Thomas, J.H. & MacCoss, M.J., 2008. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res 18(10): 1660-1669.

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. & Puhler, A., 2003. GenDB--an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31(8): 2187-2195.

Meylan, E., Burns, K., Hofmann, K., Blancheteau, V., Martinon, F., Kelliher, M. & Tschopp, J., 2004. RIP1 is an essential mediator of Toll-like receptor 3-induced NF-kappa B activation. Nat Immunol 5(5): 503-507.

Meylan, E. & Tschopp, J., 2005. The RIP kinases: crucial integrators of cellular stress. Trends

Biochem Sci 30(3): 151-159.

Miggin, S.M. & O'Neill, L.A., 2006. New insights into the regulation of TLR signaling. J Leukoc Biol 80(2): 220-226.

Muzio, M., Bosisio, D., Polentarutti, N., D'Amico, G., Stoppacciaro, A., Mancinelli, R., van't Veer, C., Penton-Rol, G., Ruco, L.P., Allavena, P. & Mantovani, A., 2000. Differential expression and regulation of toll-like receptors (TLR) in human leukocytes: selective expression of TLR3 in dendritic cells. J Immunol 164(11): 5998-6004.

Nakai, K. & Horton, P., 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24(1): 34-36.

Nielsen, P. & Krogh, A., 2005. Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics 21(24): 4322-4329.

Oganesyan, G., Saha, S.K., Guo, B., He, J.Q., Shahangian, A., Zarnegar, B., Perry, A. & Cheng, G., 2006. Critical role of TRAF3 in the Toll-like receptor-dependent and -independent antiviral response. Nature 439(7073): 208-211.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J. & Buell, C.R., 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35(Database issue): D883-887.

Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20(18): 3551-3567.

Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. & White, O., 2001. The Comprehensive Microbial Resource. Nucleic Acids Res 29(1): 123-125.

Pindado, J., Balsinde, J. & Balboa, M.A., 2007. TLR3-dependent induction of nitric oxide synthase in RAW 264.7 macrophage-like cells via a cytosolic phospholipase A2/cyclooxygenase-2 pathway. J Immunol 179(7): 4821-4828.

Pobezinskaya, Y.L., Kim, Y.S., Choksi, S., Morgan, M.J., Li, T., Liu, C. & Liu, Z., 2008. The function of TRADD in signaling through tumor necrosis factor receptor 1 and TRIF-dependent Toll-like receptors. Nat Immunol 9(9): 1047-1054.

Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M., Stabenau, A., Storey, R. & Clamp, M., 2004. The Ensembl analysis pipeline. Genome Res 14(5): 934-941.

Power, K.A., McRedmond, J.P., de Stefani, A., Gallagher, W.M. & Gaora, P.O., 2009. High-throughput proteomics detection of novel splice isoforms in human platelets. PLoS ONE 4(3): e5001.

Rappsilber, J., Ishihama, Y. & Mann, M., 2003. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem 75(3): 663-670.

Rappsilber, J., Mann, M. & Ishihama, Y., 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2(8): 1896-1906.

Ravichandran, A., Sugiyama, N., Tomita, M., Swarup, S. & Ishihama, Y., 2009. Ser/Thr/Tyr phosphoproteome analysis of pathogenic and non-pathogenic Pseudomonas species. Proteomics 9(10): 2764-2775.

Rison, S.C., Mattow, J., Jungblut, P.R. & Stoker, N.G., 2007. Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the

proteome of Mycobacterium tuberculosis. Microbiology 153(Pt 2): 521-528.

Saito, H., Oda, Y., Sato, T., Kuromitsu, J. & Ishihama, Y., 2006. Multiplexed two-dimensional liquid chromatography for MALDI and nanoelectrospray ionization mass spectrometry in proteomics. J Proteome Res 5(7): 1803-1807.

Salzberg, S.L., Delcher, A.L., Kasif, S. & White, O., 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26(2): 544-548.

Santos, S.D., Verveer, P.J. & Bastiaens, P.I., 2007. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. Nat Cell Biol 9(3): 324-330.

Sasaki, T., 1999. Current status of and future prospects for genome analysis in rice In: Shimamoto, K. (ed.) Molecular Biology of Rice Springer-Verleg, Japan.

Sato, S., Sugiyama, M., Yamamoto, M., Watanabe, Y., Kawai, T., Takeda, K. & Akira, S., 2003. Toll/IL-1 receptor domain-containing adaptor inducing IFN-beta (TRIF) associates with TNF receptor-associated factor 6 and TANK-binding kinase 1, and activates two distinct transcription factors, NF-kappa B and IFN-regulatory factor-3, in the Toll-like receptor signaling. J Immunol 171(8): 4304-4310.

Selvarajoo, K., 2006a. Discovering differential activation machinery of the Toll-like receptor 4 signaling pathways in MyD88 knockouts. FEBS Lett 580(5): 1457-1464.

Selvarajoo, K., 2006b. Decoding the Signaling Mechanism of Toll-Like Receptor 4 Pathways in Wild Type and Knockouts. In: Arjunan, S.N.V., et al. (eds.) E-Cell System: Basic Concepts and Applications Kluwer Academic Publishers - Landes Bioscience.

Selvarajoo, k., Helmy, M., Tomita, M. & Tsuchiya, M., 2008b. Inferring the mechanistic basis for the dynamic response of the MyD88-dependent and –independent pathways. Pp 110-113. 10th International Conference on Molecular Systems Biology (ICSMB 2008),

Manila, Philippines.

Selvarajoo, K., Takada, Y., Gohda, J., Helmy, M., Akira, S., Tomita, M., Tsuchiya, M., Inoue, J. & Matsuo, K., 2008a. Signaling flux redistribution at toll-like receptor pathway junctions. PLoS ONE 3(10): e3430.

Selvarajoo, K., Tomita, M. & Tsuchiya, M., 2009**.** Can complex cellular processes be governed by simple linear rules? J Bioinform Comput Biol 7(1): 243-268.

Selvarajoo, K. & Tsuchiys, M., 2007. Systematic Determination of Biological Network Topology: Nointegral Connectivity Method (NICM). Pp 449-471. In: Choi, S. (ed.) Introduction to Systems Biology The Humana Press, Inc, New Jersey.

Sevinsky, J.R., Cargile, B.J., Bunger, M.K., Meng, F., Yates, N.A., Hendrickson, R.C. & Stephenson, J.L., Jr., 2008. Whole genome searching with shotgun proteomic data: applications for genome annotation. J Proteome Res 7(1): 80-88.

Shadforth, I., Dunkley, T., Lilley, K., Crowther, D. & Bessant, C., 2005. Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds. Rapid Commun Mass Spectrom 19(22): 3363-3368.

Smalheiser, N.R., 2002. Informatics and hypothesis-driven research. EMBO Rep 3(8): 702.

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24(5): 637-644.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. & Lewis, S., 2002. The generic genome browser: a building block for a model organism system database. Genome Res 12(10): 1599-1610.

Su, X., Li, S., Meng, M., Qian, W., Xie, W., Chen, D., Zhai, Z. & Shu, H.B., 2006. TNF

receptor-associated factor-1 (TRAF1) negatively regulates Toll/IL-1 receptor domain-containing adaptor inducing IFN-beta (TRIF)-mediated signaling. Eur J Immunol 36(1): 199-206.

Sugiyama, N., Nakagami, H., Mochida, K., Daudi, A., Tomita, M., Shirasu, K. & Ishihama, Y., 2008. Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. Mol Syst Biol 4: 193.

Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y. & Tomita, M., 2003. E-Cell 2: multi-platform E-Cell simulation system. Bioinformatics 19(13): 1727-1729.

Takeda, K. & Akira, S., 2005. Toll-like receptors in innate immunity. International Immunology 17(1): 1-14.

Takeuchi, O. & Akira, S., 2008. MDA5/RIG-I and virus recognition. Curr Opin Immunol 20(1): 17-22.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S.P. & Bafna, V., 2007. Improving gene annotation using peptide mass spectrometry. Genome Res 17(2): 231-239.

Tanner, S., Shu, H., Frank, A., Wang, L.C., Zandi, E., Mumby, M., Pevzner, P.A. & Bafna, V., 2005. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77(14): 4626-4639.

Tatusov, R.L., Koonin, E.V. & Lipman, D.J., 1997. A genomic perspective on protein families. Science 278(5338): 631-637.

Tatusova, T.A. & Madden, T.L., 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174(2): 247-250.

Tyers, M. & Mann, M., 2003. From genomics to proteomics. Nature 422(6928): 193-197.

Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R. & Wishart, D.S., 2005. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res 33(Web Server issue): W455-459.

Vance, W., Arkin, A. & Ross, J., 2002. Determination of causal connectivities of species in reaction networks. Proc Natl Acad Sci U S A 99(9): 5816-5821.

Veenstra, T.D., 2006. Mass Spectrometry: The Foundation of Proteomics. In: Timothy D. Veenstra & Yates, J.R. (eds.) Proteomics for Biological Discovery John Wiley & Sons, Inc., Publication, USA.

Visintin, A., Mazzoni, A., Spitzer, J.H., Wyllie, D.H., Dower, S.K. & Segal, D.M., 2001. Regulation of Toll-like receptors in human monocytes and dendritic cells. J Immunol 166(1): 249-255.

Wang, R., Prince, J.T. & Marcotte, E.M., 2005. Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. Genome Res 15(8): 1118-1126.

Wang, X., Shi, X., Hao, B., Ge, S. & Luo, J., 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol 165(3): 937-946.

Werner, S.L., Barken, D. & Hoffmann, A., 2005. Stimulus specificity of gene expression programs determined by temporal control of IKK activity. Science 309(5742): 1857-1861.

Wiener, N., 1948. Cybernetics: Or the Control and Communication in the Animal and the Machine. Paris, France: Librairie Hermann & Cie, and Cambridge, MA: MIT Press.Cambridge, MA: MIT Press.

Wright, J.C., Sugden, D., Francis-McIntyre, S., Riba-Garcia, I., Gaskell, S.J., Grigoriev, I.V., Baker, S.E., Beynon, R.J. & Hubbard, S.J., 2009. Exploiting proteomic data for

genome annotation and gene model validation in Aspergillus niger. BMC Genomics 10: 61.

Wu, C.C., MacCoss, M.J., Howell, K.E. & Yates, J.R., 3rd, 2003. A method for the comprehensive proteomic analysis of membrane proteins. Nat Biotechnol 21(5): 532-538.

Yamamoto, M., Sato, S., Hemmi, H., Uematsu, S., Hoshino, K., Kaisho, T., Takeuchi, O., Takeda, K. & Akira, S., 2003. TRAM is specifically involved in the Toll-like receptor 4-mediated MyD88-independent signaling pathway. Nat Immunol 4(11): 1144-1150.

Yuan, Q., Ouyang, S., Wang, A., Zhu, W., Maiti, R., Lin, H., Hamilton, J., Haas, B., Sultana, R., Cheung, F., Wortman, J. & Buell, C.R., 2005. The institute for genomic research Osa1 rice genome annotation database. Plant Physiol 138(1): 18-26.

Appendix

Abbreviations

AP -1	Activator Protein-1
APS	Average Peptide Scoring
BLAST	Basic Local Alignment Search Tool
DCs	Dendritic Cells
DD	Death Domain
dsRNA	double-stranded RNA
EGF	Epidermal Growth Factor
ERK	Extracellular Signal-Regulated Kinase
EST	Expressed Sequence Tags
FDR	False Discovery Rate
IEF	Iso-Electric Focusing
IRF	Interferon Regulatory Factor
JNK	Jun N-terminal Kinase
KO	knockout
LC-MS/MS	Liquid Chromatography-Mass Spectrometry
LPS	lipopolysaccarides
MAPK	Mitogen-activated protein kinases
MDA	Melanoma-Differentiation-Associated gene
MEFs	Murine Embryonic Fibroblasts
MyD88	Myeloid Differentiation factor 88

NF-κB	Nuclear Factor-kappaB
NGF	Neuronal Growth Factor
p38	Mitogen-Activated Protein kinases 38
PAMP	Pathogen-Associated Molecular Pattern
PCR	Polymerase Chain Reaction
Poly (I:C)	Polyinosine-polycytidylic acid
PRR	pattern recognition receptors
RIG -I	Retinoic-Acid-Inducible Protein -I
SCX	Strong Cation Exchange
SFR	Signaling Flux Redistribution
TIGR DB	The Institute Of Genome Research Database
TLR	Toll-like Receptors
TNF	Tumour-Necrosis Factor
TRADD	NFRSF1A-associated via death domain
TRAF	TNF Receptor Associated Factor
TRIF	TIR domain-containing adapter-including interferon-β
WT	Wildtype