DACS-DB: AN ANNOTATION AND DISSEMINATION MODEL FOR DISEASE ASSOCIATED CYTOKINE SNPs

Sushant Bhushan

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree of
Master of Science in Bioinformatics,
Indiana University
August 2011

Accepted by the Faculty of Indiana University,
in partial fulfillment of the requirements for the degree of Master of Science
in Bioinformatics

Master's Thesis Committee

Dr. Narayanan B. Perumal, PhD, Chair

Dr. Malika Mahoui, PhD

Dr. Todd Skarr, PhD

Dedicated to my Family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES:

ACKNOWLEDGEMENTS

I would like to take the opportunity to acknowledge some of the people who made my graduate study a memorable experience and made this thesis possible. Words are really inadequate to express my gratitude and regards towards my advisor Dr. Narayanan B. Perumal. His hard work ethic, encouragement, insight and vivid descriptions of concepts, attention to details, valuable advice, motivation, supervision, valuable input, and corrections of write up inspired me to complete this work successfully.

I would like thank to Dr. Malika Mahoui and Dr. Todd Skaar for their time, encouragement, advice for improvement and critical evaluation. I greatly appreciate their efforts in reviewing my work. I appreciate their kind analysis, valuable input and correction of the thesis work and thesis writing which helped a lot.

I owe my sincere thanks to Indiana University for providing the prestigious Dean's award during my Master's program. I would like to extend my sincere thanks to Dr. Mathew Palakal, Mark McCreary, Dr. Pedro Romero, and other faculty and staff of the department of Bioinformatics for their continuous support throughout the Master's program.

Above all, I express my deep gratitude and affection to the God, my parents and family members whose persistent love, warmth, cooperation and encouragement were the main reason for the successful completion of my thesis. I would like to extend my sincere thanks to Myron Snelson for doing a project in machine learning and his help in classification of SNPs into disease and non-disease classes and his help with Python programming and execution of an SVM tool. I would like to thank my friends Rahul, Anusha, Rini, Satish, Prem, Gokhul, Prudhvi,

Kanishka, Anand, Pratheek and Namritha for their support and all the fun we have had in the last two and half years.

.

ABSTRACT

Cytokines mediate crucial functions in innate and adaptive immunity. They play valuable roles in immune cell growth and lineage specification, and are associated with various disease pathologies. A large number of low, medium and high throughput studies have implicated association of single nucleotide polymorphisms (SNPs) in cytokine genes with diseases. A preponderance of such experiments have not shown any causality of an identified SNP to the associated disease. Instead, they have identified statistically significant SNP-disease associations; hence, it is likely that some of these cytokine gene variants may directly or indirectly cause the disease phenotype(s). To fill this knowledge gap and derive study parameters for cytokine SNP-disease causality relationships, we have designed and developed the Disease Associated Cytokine SNP Database (DACS-DB). DACS-DB has data on 456 cytokine genes, approximately 61,000 SNPs, and 891 SNP-associated diseases. In DACS-DB, among other attributes, we present functional annotation, and heterozygosity allele frequency for the SNPs, and literature-validated SNP association for diseases. Users of the DB can run queries such as the ones to find disease-associated SNPs in a cytokine gene, and all the SNPs involved in a disease. We have developed a web front end (available at http://www.iupui.edu/~cytosnp) to disseminate this information for immunologists, biomedical researchers, and other interested biological researchers. Since there is no such comprehensive collection of disease associated cytokine SNPs, this DB will be vital to understanding the role of cytokine SNPs as markers in disease, and, more importantly, in causality to disease thus helping to identify drug targets for common inflammatory diseases.

Due to the presence of rich annotations, the DACS-DB can be a good source for building a tool for the prediction of the “disease association potential (DAP)” of a given SNP. In a preliminary effort to devise such a methodology for DAP prediction, we have applied a support vector machine (SVM) to classify SNPs. Employing the SNP attributes of function class, heterozygosity value, and heterozygosity standard error, 864 SNPs were classified into two classes, “disease” and “non-disease”. The SVM returned a

classification of these SNPs into the disease and non-disease classes with an accuracy of 74%. By modifying various SNP and disease attributes in the training data sets, such a predictive algorithm can be extrapolated to identify potential disease associated SNPs among newly sequenced cytokine variations. In the long run, this approach can provide a means for future gene variation based therapeutic regimens.

CHAPTER ONE: INTRODUCTION

1.1 Cytokines

Cytokine molecules regulate the innate and adaptive response of the body [1, 36]. They mediate lymphopoiesis, myelogenesis and stem cell and tissue differentiation. They play roles in growth and development of the body. Cytokines are small proteinaceous molecules with autocrine, paracrine and endocrine mode of action.

Types of Cytokines

Cytokines are divided into the following six classes of molecules:

i.) Interleukins

ii.) Interferons

iii.) Chemokines

iv.) Tumor necrosis factor family

v.) TGF beta family

vi.) Growth hormones

Mechanism of action of Cytokines

Cytokines act on the invading pathogens by activation of intermediary molecules that activate T-cells and B-cells [2]. They also regulate the production of antibodies by B-cells and cell apoptosis mediated by monocytes, macrophages and natural killer cells. They also mediate the early line of defense of the body and also play roles in the adaptive immune response of the body. Cytokines mediate the inflammation occurring in the body and also play role in the growth and differentiation of cells.

Classification of Cytokines

Cytokines are classified as proinflammatory and anti-inflammatory. Proinflammatory cytokines cause and mediate the inflammation in the body while anti-inflammatory cytokines mediate the curbing of inflammation in the body. All the major diseases including inflammatory and auto-immune diseases are associated with cytokines. Diseases ranging from cancers, cardiovascular diseases, diabetes, skin disorders, kidney disorders, psychological disorders (mood disorders, depression, ADHD, bipolar diseases etc.) , Parkinson disease, Alzheimer’s disease, inflammatory disorders such as systemic lupus erythematosus, multiple sclerosis and Crohn’s disease have one or many cytokines associated with them.

1.2 SNPs

Single nucleotide polymorphism

Single nucleotide polymorphism is a single nucleotide variation present in the genome of an organism and found to be present in at least 1% of the population. This variation is due to the presence of different nucleotides in different allelic positions. It can be bi-allelic, tri allelic or tetra-allelic but bi-allelic is the most commonly observed variation [4]. This allelic variation is present in similar nucleotide sequence and hence can act as marker.

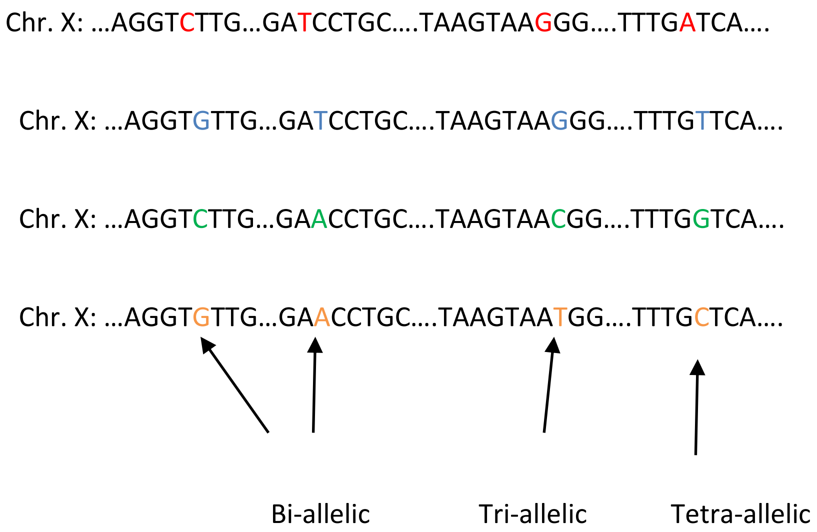

Chr. X: …AGGTCTTG…GATCCTGC….TAAGTAAGGG….TTTGATCA….
Chr. X: …AGGTGTTG…GATCCTGC….TAAGTAAGGG….TTTGTTCA….
Chr. X: …AGGTCTTG…GAACCTGC….TAAGTAACGG….TTTGGTCA….
Chr. X: …AGGTGTTG…GAACCTGC….TAAGTAATGG….TTTGCTCA….
Bi-allelic
Tri-allelic
Tetra-allelic

Fig.1: In the four different individuals, in same region of the chromosome X, four different SNPs are shown to be present at specific positions among similar nucleotide bases. The first two SNPs are bi-allelic(G,C),the third is tri-allelic(G,C,T) and the fourth SNP is tetra-allelic(A,T,G,C).

"When two random chromosomes are compared, they differ at ~1⁄1000 nucleotides [6]. When all chromosomes from 40 individuals are screened, about 17 million SNPs are expected to be found, out of the 3 billion bases in human DNA. Only a small proportion of these SNPs are expected to be in coding regions, as coding regions are ~1% of the genome and are less likely to have SNPs. Thus, the number of coding SNPs is estimated to be ~500,000, which is an average of about 6 SNPs per gene" [3].

Comparison to other markers

Restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), random amplification of polymorphic DNA (RAPD), single sequence repeat (SSR), and expressed sequence tags (EST) are the other markers that can be used to characterize a genome. Compared to these markers, SNPs are more abundant, less costly and do not require assays for their determination. AFLP and SSR are poly-allelic while SNP is bi-allelic causing it to have less fidelity in inheritance. Despite being bi-allelic in nature, SNPs are better biomarkers, as they are present in large numbers in the genome [1, 36].

1.3 Literature survey on cytokines, SNPs and diseases

1.3.1 Cytokines

Role and Significance of cytokines

Cytokines act on the invading pathogens by activation of intermediary molecules that activate T-cells and B-cells [1,36]. They also regulate the production of antibodies by B-cells and cell apoptosis mediated by monocytes, macrophages and natural killer cells. Cytokine molecules mediate the early line of defence of the body and also play a role in adaptive immune response of the body. Cytokines mediate the inflammation occurring in the body and also play role in the growth and differentiation of cells.

Cytokines are classified as Proinflammatory cytokines and anti-inflammatory cytokines. Proinflammatory cytokines cause and mediate the inflammation in the body [8] while anti-inflammatory cytokines mediate the curbing of inflammation in the body [9]. All the major diseases including inflammatory and auto-immune diseases are associated with cytokines ranging from cancers, cardiovascular diseases, diabetes, kidney disorders, skin disorders, psychotic disorders (mood disorders, depression, ADHD, bipolar diseases etc.) , Parkinson disease, Alzheimer’s disease and inflammatory disorders such as Systemic lupus erythematosus, Multiple sclerosis, Crohn’s disease[2].

Cytokine role in bone formation

Cytokines play a major role in bone formation. IL-11 and IL-6 mediates the generation of osteoclast cells while TNF and IL-1 promote the generation of Osteoclast cells by promoting production of IL-11 and IL-6 cytokines which promote bone loss. IL-6 in combination with IL-3 has been observed to promote osteoclastogenic effects in mice in vitro, while in combination with IL-1 has osteoclastogenic effect in-vivo. Interleukin-6 has a pathogenetic role in the abnormal bone resorption associated with multiple myeloma, Paget's disease, rheumatoid arthritis, and Gorham–Stout disease (also known as vanishing- or disappearing-bone disease) [10].

Cytokines in diseases

IFN-β plays important role in treatment of multiple sclerosis but its role is not completely understood; IFN-β might act as an antagonist of endogenous IL-4 and, particularly, IFN-γ; the latter is known to exacerbate multiple sclerosis. IFN-β influences the function of the blood–brain barrier by inhibiting cell adhesion, cell migration and metalloproteinase activity. It also induces a shift of the cytokine profile to an anti-inflammatory phenotype (e.g. TGF-β and IL-10). However, other evidence indicates that IFN-β might also up regulate some pro-inflammatory responses, such as IL-12, chemokines and their receptors. .IL-1 was the first cytokine detected by bioassay in synovial fluid of patients with Rheumatoid arthritis, followed by the demonstration of TNF, IL-6, IL-2, GM-CSF and other, chiefly pro-inflammatory, cytokines in cultures of rheumatoid arthritis synovium. IL-1 is regulated by TNF, GM-CSF and immune complexes in this disease. Anti-TNF-induced reductions have been noted in levels of IL-6, IL-1, VEGF, IL-8 and other chemokines in rheumatoid arthritis synovium [11,12,14].

Th1 cells regulate cell mediated immunity and secrete IL-1, IL-12, IFN$_\gamma$ and TNFα. Th2 cells mediate humoral immunity and secrete IL-4, IL-5, IL-6, IL-10 and IL-13. IL-4 secreted by TH2 suppresses the activity of Th1 cells while IFN$_\gamma$ secreted by Th1 cells suppresses the activity of Th2 cells. Crohn’s disease is considered Th1 disease while ulcerative colitis is considered having the characteristics of Th2 cells. “Models of Inflammatory bowel disease suggest that suppression of activated T lymphocytes may be defective in the disease. Lack of IL-10 causes inflammatory bowel diseases but has increased amount of IL-1, TNFα, IL-6 and IFN$_\gamma$ resulting in suppression of activation of anti-inflammatory cytokines and causing chronic inflammation”. TNFα mediates the production of pro-inflammatory cytokines and alters the behavior of cells that are intimately involved in immune response[11].

Increased synthesis of prostaglandins is a major feature of both acute and chronic inflammatory responses and is stimulated by a number of cytokines that are present at inflammatory sites, including interleukin 1 (IL-1), tumor necrosis factor (TNF), transforming growth factor β (TGF-β) and platelet-derived growth factor (PDGF). IL-l and IL-6 are involved in this hyperalgesic response, and that production of these cytokines is induced by bradykinin and TNF-α. “Injection of antibodies to IL-1β or IL-6 attenuates fever and thermogenesis induced by endotoxin, It has been demonstrated that certain cytokines (IL-6 and IL-8) elicit fever and thermogenesis via release of Corticotrophin releasing hormone (CRH), whereas others (IL-1α and TNF-a) act independently of Corticotrophin releasing hormone (CRH)".

Severe combined immunodeficiency disease is implicated to be associated with IL-2$R_γ$ which is component of receptors for IL-4, IL-7, IL-9, IL-15 and IL-21. Th 1 -type responses appear to be involved in organ specific autoimmunity, in contact dermatitis, and in some chronic inflammatory disorders of unknown etiology. In contrast, in genetically predisposed hosts, Th2-type responses against common environmental allergens are responsible for triggering of allergic atopic disorders. Altered profiles of lymphokine production may account for immune dysfunctions in some primary or acquired immunodeficiency syndromes. Th l -type responses appear to be involved in organ specific autoimmunity, in contact dermatitis, and in some chronic inflammatory disorders of unknown etiology. In contrast, in genetically predisposed hosts, Th2-type responses against common environmental allergens are responsible for triggering of allergic atopic disorders. Altered profiles of lymphokine production may account for immune dysfunctions in some primary or acquired immunodeficiency syndromes. IFN-α, IL-10, IL-12, and/or transforming growth factor TGF-β produced by macrophages and B cells have important effects on CD4+ subset maturation. All but IL-10 preferentially induces Th 1 expansion in various systems, whereas IL-10 was implicated in Th2 expansion when spleen cells, as opposed to B cells, were used as Antigen presenting cells (APC). IFN-y promotes differentiation to Th I cells, both in vitro and in vivo. Interestingly, T cell-independent IFN-y production by natural killer (NK) cells in response to IL-12 has recently been reported. Thus, IL-12 may be critical as natural initiator for the development of Th 1 responses[16,17,18,19,20]. ”Most striking is the requirement for IL-4 for maturation of naive Th cells into Th2 cells. Omenn's syndrome is a rare and severe combined immunodeficiency, characterized by hyper eosinophilia and increased serum levels of IgE. IL-4-induced IgE molecules are pathogenetically involved in atopy-associated diseases, such as allergic asthma, allergic rhinitis, and atopic dermatitis. Hashimoto's thyroiditis or Graves' disease has a clear-cut TH1 Iymphokine profile with production of high TNF-α and IFN-y concentrations. The levels of IL-6, IL-2, TNF-α, and IFN-y have been found to be elevated in the serum of patients with active systemic lupus erythematosus (SLE) or with particular forms of the disease. Active rheumatoid arthritis contains significant amounts of macrophage and/or fibroblast products, including IL- l ,I L-6,I L-8,T NF-α,"[11,16,18]. In contrast to pro-inflammatory cytokines, T cell-derived lymphokine, such as IL-2, IL-3, IL-4, IFN-y and TNF-β, are present in synovial fluid of rheumatoid arthritis patients in low concentrations. Crohn’s Disease patients produced IL-2 and IFN-y, but not IL-4 and IL-5. Increased production of IL-2, TNF-α, and/or IFN-y by CD4 + T cells present in lungs and lymph nodes have been found in active sarcoidosis. GM-CSF, TNF-α, and IFN-y are associated with tumors. IL-1β and TNF influence the severity of possibly by persistent activation of NF-KB as observed in case of rheumatoid arthritis. IL-1, IL-6, TNF are observed with microglia which is constantly observed in neural plaques of rheumatoid arthritis while IL-1, IL-6, TNF are observed in astrocytes. IFN-$_{\gamma}$ and TNF play disease promoting roles multiple sclerosis while anti-inflammatory cytokines such as IL-10 and TGF-β are disease down regulating. “IL-2, IL-12 activates TH1 while IL-2, IFN-$_{\gamma}$, TNF-β mediates cell mediated immunity. IL-4, IL-13 activates TH2 cells while IL-4, IL-5, IL6, IL-10, 1L-13 released by TH2 cells activates humoral immunity. IL-2, IL-3, IL-4, IL-5, IL-6, IL-10, GM-CSF, IFN-$_{\gamma}$, TNF-β down regulates TH1 cells. TH1 plays role in viral infections and graft versus host reactions while Th2 plays role in immune memory, allergy and parasite infections, proinflammatory cytokines such as TNF, IL-1, IL-6, GM-CSF, and chemokines such as IL-8 are abundant in all patients rheumatoid arthritis. This is compensated to some degree by the increased production of anti-inflammatory cytokines such as IL-10 and TGF-β and cytokine inhibitors such as IL-1RA and soluble TNF-R. TNF-α, TNF-β, IL-6 activate HIV directly through the induction of NF-KB and IL-6 through the synergistic effect with TNF-α. IL-2 activated by IL-1 activates T-cells in an autocrine manner. Activated T-cells produce Cytokines such as TNF-α, TNF-β, IL-3, IL-4, GM-CSF and IL-6 all of which, in turn activates HIV replication in T-lymphocytes and macrophages in an autocrine and paracrine manner. IL-6 is a growth factor for AIDS Kaposi sarcoma cells while IL-1 and TNF-α are involved in pathogenic mechanism of Kaposi sarcoma. TNF-*α*, IL-1*β*, viruses [such as CMV (cytomegalovirus)] and bacterial pathogens (such as *Chlamydia pneumoniae* and *Helicobacter pylori*) can activate NF-*κ*B, which binds to specific sites on the promoter regions of target genes. This, in turn, modulates the endothelial synthesis of proinflammatory cytokines (IL-1, IL-6 and TNF-*α*) and chemokines, IL-8, MCP-1 and RANTES (**r**egulated upon **a**ctivation, **n**ormal **T**-cell **e**xpressed and **s**ecreted). Cytokine stem cell factor GM-CSF and erythropoietin relaxed, whereas TNF-*α*, IL-6 and IL-10 induced contraction of, human arterial segments and thus affecting the heart condition. the proinflammatory cytokines TNF-*α* and IL-1*β* (but not IL-6) induced transient and reversible endothelial dysfunction in humans. High transcardiac levels of soluble TNF-*α*R1 (TNF-*α*-receptor p55) seem to be protective of endothelial function (probably by inactivating circulating TNF-*α*), high IL-6 and soluble IL-2R (IL-2-receptor) levels were associated with impaired microvascular function. IL-6 is the main hepatic stimulus for CRP. Elevated plasma levels of IL-6 and TNF-*α* were detected consistently in patients with stable or unstable angina and myocardial infarction. In the atherosclerotic plaque, cytokines are released from macrophages, dendritic cells, T-cells, ECs and smooth muscle cells (IL-1β, TNF-*α*, IL-6, IL-8 andMCP-1). The unstable plaque is characterized by infiltrating Th1 cells, producing IFN-*γ* , IL-2, IL-6 and TNF-*α.* In aged vessels, expression of TNF-*α*, IL-1*β*, IL-6, IL-6Rα (IL-6-receptor α) and IL-17 genes was significantly increased compared with young vessels. Anti-inflammatory cytokines exerting inhibitory effects on vascular cells include TGF-β, IL-10, TGF-β and IL-1ra with acute coronary syndromes in angina patients. proinflammatory cytokines (IL-1, IL-6 and TNF-α) can exert negative inotropic effects and, therefore, directly modulate cardiac contractility. Obesity also leads to a proinflammatory and prothrombotic state that potentiates atherosclerosis. Recent findings imply a role for fat-derived 'adipokines', including TNF-α, IL-1β, IL-6, IL-8, IL-10, TGF-β and adiponectin, as pathogenic contributors or protective factors. Angiogenesis (i.e. the generation of new capillary blood vessels from pre-existing vasculature) is potentiated by VEGF, FGF, PDGF and TGF-β and via the chemokines MCP-1 and MIP (macrophage inflammatory protein). Cleaved fragments of both receptor types, also known as TNF-binding proteins (TNF-BPs), have been detected in the urine and serum of patients with a variety of diseases, including cancer, AIDS, and sepsis. TNF is implicated in diseases as septic shock, cancer, rheumatoid arthritis, malaria, and other affliction. TNF has also been implicated in the pathobiology of cancer cachexia, but its role in this metabolic disease is less well defined than in septic shock syndrome. TNF production is increased in macrophages obtained from cancer patients. Prolonged exposure of animals to TNF causes cachexia with typical losses of protein, lipids, and red blood cell mass. Also, in cancer cachexia, TNF-induced cachexia also causes insulin resistance and derangements of glcuose metabolism." [11-20]

1.3.2 SNPs

SNPs and their significance

Single nucleotide polymorphism marks the difference of nucleotide composition occurring in individuals marked by a difference in nucleotide at a specific location of genome of individuals in a specific population and also the variation observed to be found out in at least 1% of the population. SNPs can be causally associated with disease if the variation can result in alteration of amino acid composition resulting in change in protein expression and hence causally been associated with disease phenotype; these SNPs (missense mutations) are called non-synonymous SNPs. If the nucleotide variation does not affect the expression resultant from the nucleotide variation, they are called as synonymous SNPs. SNPs can act as the biomarker if they are observed in haplotypes of a particular population for a specific phenotypic expression and they can represent the genetic loci as marker for the particular phenotypic expression. It can also represent selecting other SNPs as biomarker if the other SNPs are present in linkage disequilibrium with the SNP under study. These SNPs can act as the biomarker for the particular phenotype and can represent the haplotype for the particular population group. SNPs, hence can be a better genetic markers compared to the genes if they can represent the genetic loci with high statistical significance for particular phenotypic trait [21, 29, 30].

SNP role in causality of disease

SNPs can be responsible for causality of a disease, if the amino acid variation caused by them result in change in protein expression and also change in the phenotypic expression which can result in a diseased state. This non-synonymous SNP change might be resultant because of mutation brought about by the environment factors or the inherent changes in the nucleotide expression resulting because of difference in inheritance patterns. If the result of this variation is non-synonymous SNP, it can be causally associated with diseases, if the variation is synonymous SNP; it has less possibility of being associated with disease. If the non-synonymous SNP is present in exonic region, it has possibility of being associated with disease, compared to non-synonymous SNP present in the 5’ UTR, 3’ UTR, near-gene-5, near-gene- 3 and intron. The same can be true for synonymous SNPs present in the exonic region, or non-exonic region. “The SNPs which are present in high linkage disequilibrium to the synonymous SNP can be causally associated with disease if they have any cumulative effect on the genetic expression of the genetic element i.e. the change in the haplotypic expression can bring about any expression change across the species and bring about change across the species; such a variation can be considered to be causally responsible for change in the genetic expression and hence can be responsible for a causality of pandemic disease ". The causal role of SNPs can be crucial for understanding the role of genetic cause of disease at a single nucleotide level which can be crucial in understanding the effect of mutational changes on the disease causes, since these mutations when inherited and present in at least 1% of the population are classified as single nucleotide polymorphism. Most of the SNPs are determined to be associated to diseases statistically and hence they can be said to be stochastically associate to diseases, hence the causal association can be established only by biological validation.

SNPs significance as biomarker for diseases

Before the wide scale study of SNPs, most of the previously studied markers were cytogenetic markers, dependent upon the band patterns and the markers were a repeated sequence of nucleotides and they are distributed across the genome at regular interval, present in the genome with a unique expression pattern .This causes of decrease in expression detection of the biomarkers and also most of the times these biomarkers are tri allelic and tetra allelic, hence their expression biologically does not contain any significance; hence SNPs which are bi-allelic are more prominent and effective as biomarker. SNPs which are selected to act as biomarker for the gene are usually present in the genome at high linkage disequilibrium with other SNPs and hence these groups of SNPs as

biomarker can represent the haplotype of the population group. Most of the times synonymous SNPs can be represented as biomarkers but the synonymous SNPs represent the haplotype instead of representing it as a single nucleotide variation. Non-synonymous SNPs can be represented as the biomarkers as they have marked effect on the amino acid expression and hence phenotypic expression and hence they can represent the genetic elements which can affect the phenotypic expression associated with diseases and can be the biomarker. SNPs most of the time can be suitable marker for the diseases and can suitably represent the genetic element for the disease and hence SNPs can be a suitable target for treatment of diseases. SNPs as biomarker can be good source for representing the drug targets and drug development. SNPs as biomarker can also represent the variations occurring in the genome because they can be representative of the genes as they have specific chromosomal location with in genes [21, 29, 30, and 31].

Importance of SNPs as biomarkers

SNPs as biomarkers can be a specific representation of a gene and they can reduce the genotyping cost associated with analyzing a complete gene compared to a single nucleotide. Single nucleotides polymorphism when compared with a gene, is easy to genotype and its genetic association can be easily determined, as its statistical significance can be easily established. SNPs as biomarkers can represent the genetic element associated with disease and hence can be very important in understanding the genetic inheritance of the disease; They can also help understand the evolutionary lineage of a gene's orthology and paralogy associated with genetic evolution. This can also help understand the genetic drift and genetic bottleneck phenomena and hence can be crucial in understanding the genetic basis of disease being associated with genetic evolution and to be more specific whether they can be associated with genetic evolution or not. SNPs as genetic markers can represent the genetic basis of disease causality and hence can be crucial in personalized medicine and detecting diseases at an early stage or even with the possibility to predict the diseases at early stages; this can also help in predicting the diseases for an individual whose genetic make-up can be determined by the knowledge of SNPs present in the genome and hence suggesting propensity of occurrence of a disease on the basis of SNP composition of the individual in the future. Individuals with specific genetic make-up specifically SNP composition for any individual for a disease can also be a crucial marker for individuals with similar genetic composition and hence can pave the path for tailor made medicine and personalized medicines also considering the individuals genetic make-up in mind. SNPs as biomarkers can be significant in understanding whether the genes associated with linkage disequilibrium as a haplotype can be associated with disease and whether they can be representative for other individuals with similar nucleotide composition can be associated with diseases or not. Single nucleotide polymorphism can also represent the disease association potential for individuals with particular SNP missing and whether it can affect the targeting drugs or not. SNPs can also affect the drug targets and the impact of the drug action on the specific drug targets and whether varying the SNP composition can affect the drug activity on the individual or not. SNPs with the particular SNP composition which can be varied to a particular SNP composition can be associated with disease or not with similar SNP composition or similar haplotype pattern [21, 29, 30, and 31].

1.3.3 Diseases

Cure for diseases

Diseases are the disturbances in the natural state of equilibrium of cellular process and proper functioning of cellular state and cellular mechanism resulting in healthy state of body. Diseases are caused by external factors such as microbes, harsh conditions or imbalance in the cellular defence mechanism or disturbance in any of the inherent cellular functions. Diseases can be cured by the administration of drugs, proper diet, exercises etc. Any drug which can restore the immune system and maintains the cellular mechanism to fight against the invading pathogens can be a cure for the disease. Most of the drugs determined are based on the chemical action of drugs on the cellular mechanism or immune system to maintain and restore the cellular mechanism to maintain the body equilibrium. Drugs designed for specific diseases provide treatment for the disease. There is wide variety of drugs available for treatment of diseases e.g. Allopathic, homeopathic, Ayurvedic. Most of the drugs designed for diseases are based on the chemical methods of synthesis of drugs based on the knowledge of disease and chemical nature of disease. Most of the diseases are dependent on drugs treatment and the drugs are synthesized on the basis of previous knowledge of the diseases and the chemical composition of drugs which can act on the cause of disease and can provide cures for the treatment of disease. Most of the drugs synthesized today are chemically synthesized which are not specific for the individuals or the treatment of disease has lots of deleterious effects or allergic reactions. Drugs with genetic knowledge of the disease can help provide treatment for disease at the earlier stage. This knowledge can provide specific knowledge for the treatment of the disease as the exact location of drug target is known which can help in making personalized medicine a prominent approach for the treatment for diseases in the near future. There are lots of biological markers available for the diagnostics of the genetic cause of disease but SNPs as are present in large numbers and they have a specific location with in genome seems to be the solution for the treatment of the diseases, as they can, act as the better biomarkers for representing genetic causes of diseases when compared to the other biomarkers [22-28].

Importance of genetic treatment compared to traditional treatment

SNPs can act as genetic marker which can provide information on genetic elements which can be target for the action of drugs. They can provide a solution for the design of drugs as genetic the knowledge about specific genetic element will be known. Drugs as treatment for the cure of diseases are mostly dependent on chemical composition but introduction of genetic treatment can provide the treatment of disease at an earlier stage as genetic make-up of individual can provide information on predisposition of individual to diseases. In the future, most of the diseases can be treated by the genetic treatment and also the knowledge of SNPs can increase the understanding of drug-target interaction. Drugs designed with keeping SNPs in mind can improve the drug-gene interaction as it can provide the information on specific genetic target for the drugs. This can improve the drug target specificity and improve upon the interaction. Drugs designed with the knowledge of SNPs be more specific and efficacious and hence can improve upon the interaction drugs and drug targets. This genetic knowledge can provide the better models for designing drug targets for treatment of diseases [31].

Current status of genetic treatment

SNPs are finding a very important role in pharmacogenetics study as they can provide genetic counseling and provide avenue for personalized medicine. Drugs based on SNP analysis are becoming prevalent in the market and doctors are prescribing SNP analysis for the patients. SNPs with the pharmacogenetics analysis can provide methods for the treatment of the diseases with the information about the genetic component for the individual; it can help in explaining the variations which can lead to disease causality to the individuals. Several companies provide information on the SNP data which result in the association of diseases to drugs and which can provide ample information for designing a personalized drug suitable for an individual's genetic compositions. There are few companies present which provide information about genetic composition and SNP details of individuals e.g. 23andMe, Illumina, 454 Life sciences, Affymetrix etc. The current state of genetic treatment with SNP is at the initial stage and only involves diagnostic and providing SNP profiling and genetic composition of individuals or SNPs associated to groups of diseases. A lot of work is needed to provide genetic treatment and provide drugs and personalized care at this stage. It will need some time for personalized medicine and tailor made drugs to come into market [31, 32, 34, and 35].

Future of genetic treatment of diseases

Genetic treatment and pharmacogenetics can be the cure for future as the treatment provided by them will be more effective and more specific, and this specificity may cause less side reactions and allergy.

They can provide cures for the treatment of diseases at the genetic level which will mean premonition of disease as the genetic composition can suggest the propensity of an individual to develop a particular disease according to the presence of specific SNP composition and presence of SNP in the individual genome and the likelihood of a person to develop the disease. Individual genetic makeup can be a marker for the drugs to be administered according the genetic profile or the genetic profile of the individual race or particular population group or according to the closest genetic makeup SNP profile drugs which can have least adverse drug reaction and allergic reaction and specific, effective, more reliable treatment for the disease. This knowledge can also help increase the role of understanding the role of cytokines and SNPs in the causality of disease and hence making the genetic treatment before hand to prevent the disease from occurring in the future. It can improve the pharmacogenetics, pharmacokinetics, and genetics behind the cure of diseases and cure of disease with better understanding of drug, its target and its development with the knowledge of exact cause of disease. This can be vital in understanding and designing the drug and development of drugs in the future with more secure, healthy, personalized, target specific, efficacious and much improved drug developed for the cure of diseases and treating disease can be easier and preventing mortality because of lack of cure for diseases or because of adverse side reactions resulting from action of drugs allergic response by the body [31-38].

CHAPTER TWO: BACKGROUND

2.1 Current state of affairs

Large numbers of genome wide association studies have established the association of SNPs to diseases. With the decreasing cost of genotyping methodology the number of available SNPs is increasing. Currently, in the build 132 of dbSNP database, there is a total of 12,212,318 SNP data present for human beings. Currently there are approximately ten SNP databases available online. They are dbSNP, dB GAP, Pharm GKB, SNP500Cancer, HGVBaseG2P, JSNP, F-SNP, HAP MAP, SNP FUNCTIONAL PORTAL, SPSMART, FESD, ALFRED, CGAP SNP Index, Forensic SNP index, Gene Viewer, GeneSNP, Genome Variation Server, GWAScentral, PhenCode, SeattleSNP, SNAP. Out of these databases F-SNP, HAPMAP, SNP500Cancer have information on the phenotype association data and disease information.

SNP databases with phenotype, disease information

F-SNP

This database contains information on functional effects of SNP obtained from 16 different bioinformatics tools [32]. Functional effects are predicted and indicated at the splicing, transcriptional, translational and post-translational level. F-SNP returns information on gene synonyms, reference DB, chromosomal location, description, gene type, functional SNP, chromosomal location and region. It has a total number of 115,356 SNP data.

HAP MAP

SNP Information retrieved from TSC (The SNP Consortium) includes observed alleles, submitting lab, flanking sequences, genomic location, reported allele frequencies, population type, and protocol [33].

HAP MAP has SNP karyogram information, which provides the option to locate the disease location on all the 23 chromosomes.

SNP500Cancer

It contains information on SNP data involved in cancers and act as resource for identification and characterization of genetic variation in genes [34]. SNP data available with SNP500Cancer is a list of genes/SNPs, validated sequence results, genotype of each of 102 samples, summary of genotype data by sub-populations, primers, probes and conditions for genotyping assays, links to assay ordering information, presentation of haplotypes for all genes, GO (gene ontology) pathways, list of genes currently under analysis.

HGVBaseG2P

The Human Genome Variation database of Genotype-to-Phenotype information (HGVbaseG2P) provides a centralized compilation of summary level findings from genetic association studies, both large and small [35]. It has actively gathered datasets from public domain projects, and encourages direct data submission from the community. It has information about SNPs associated to the phenotype expression studies as it has been found out from several genotype-phenotype association studies.

SNPs3D

This database has information on the genes and their associated SNPs and diseases [37]. It tries to predict the effect of SNP on the structure and profile of the SNP. SNP with negative value is considered to be damaging. It predicts the SNP effect by structure and profile. It shows the effect of SNP on the molecular effect of the protein structure and also change in the protein structure. The disease information is for the genes but there is no information providing association of SNPs to diseases.

Genetic Association Database

GAD database finds the genetic association of gene polymorphism to complex diseases and disorders [38]. The disease page contains information on the genes associated, genetic location and disease class. The gene page contains genes, different references, OMIM Id, genetic location, disease class, and phenotype information while the polymorphism page provide information on disease associated, phenotype information, SNP Id and genetic information and different references.

Disease association of SNP databases

All the above described databases available online do not have disease information. These databases only provide information about the SNPs and their attribute information. They are specific to population groups studied except for HAPMAP which has SNP data available for 11 populations. HGVBase2 database provides information about genotype to phenotype association, while SNP500Cancer database provides information about association of SNPs to cancers while F-SNP database provides information about functional effects of SNPs at transcriptional level, splicing level, translational and post-translational level.

Current status of Cytokine study and disease association

Within the last 15 years, the Bidwell group had collected and published information about cytokine SNP and disease association. The database was available online at http://www.pam.bris.ac.uk/services/GAI/cytokine4.htm but is not available online now. The database contained information on the role of cytokines polymorphism in diseases. Later on three supplements of the paper got published in which they added further cytokine polymorphism associated to diseases data to the database. In this study Bidwell et. al.[36] showed the effect of allele polymorphism in the cytokine genes association to diseases as found in the in vitro studies. They examined the expression studies of cytokines in vivo whether it decreased or increased by the allele polymorphism. The study tried to find out whether the effect of cytokine polymorphism affects in vivo expression that can be relevant to disease association (J. Bidwell, 1999)[36]. Most of the cytokine polymorphism had suggested the association of cytokine genes to diseases. But, as suggested by the example of TNFα and LTA (alias TNFβ) the expression level is associated with the cytokine expression; and this association combined with linkage disequilibrium study can suggest the association of haplotypes to HLA genes which can be associated with diseases; this highlights the effect of cytokine polymorphism on complex diseases.

Knowledge Gap

The cytokine polymorphism study performed by the Bidwell et al group had information on polymorphism information of 74 cytokine genes. The database contained information on association of polymorphism (allelic variation) of cytokines to diseases. The study contained expression data of cytokine genes and how the expression in vitro varied for polymorphism. The database did not have information on other important attributes such as chromosome and chromosomal location of alleles, function class and other annotation information of the SNPs. Other available SNP databases such as F-SNP provide information about functional annotation of the SNP, while HapMap database provide information about the distribution of SNP across 11 population groups and also the chromosomal location of SNP associated with diseases across the 23 chromosomes. SNP500Cancer database provides information about the SNPs associated with cancers while HGVBaseG2P database provides information about association of genotypes to phenotypes. But none of these databases provides comprehensive information on association of SNPs to diseases for cytokine genes.

Need for DACS-DB database

DACS-DB database fills this knowledge gap. It has almost combined information of the six databases (F-SNP, HAPMAP, SNP-500 Cancer, HGVBaseG2P, SNPs3D, GAD) mentioned above. It provides information on association of Cytokines SNPs to diseases with vital information of dbSNP identifier and PMID citation to verify the authenticity of SNPId and disease association information to cytokine genes. DACS- DB also crucially provide important information on attributes of SNPs; chromosome, chromosomal location, function class information and also sequence information is very crucial information for drug target identification while including heterozygosity value, heterozygosity standard error information can provide along with other attribute information can provide annotation information which can be important in understanding the causal association of SNPs to diseases. Since, non-synonymous and synonymous SNPs have important information for the association of SNPs to diseases, along with heterozygosity value can critically provide association of diseases to SNPs; as SNPs present in a coding region have higher possibility of getting associated with diseases as well as SNPs present in a non-coding region, regulatory region, promoter region or presence in enhancer or silencer region can play a role in the expression of the synonymous SNP in the causality of diseases as it can affect the gene expression and gene regulation.

Goals and objectives for thesis

Goal

The goal of the thesis was to find the causal relation of cytokine SNPs to diseases and to find the cytokine SNPs associated with disease data from PubMed literature, curate and annotate the data with

SNP and cytokine data and make it available over the web as a database with suitable queries, so that, it can be accessed and be retrieved easily by the user.

Objective of the thesis

The objective of the thesis was to:

- Retrieve the SNP data from dbSNP with specific attributes which can meaningfully assign value in predicting disease association potential for SNPs.
- To retrieve cytokine SNPs associated with diseases as annotated in the relevant PubMed literature.
- To create a relational database to represent association between Cytokine, SNPs and disease tables.
- Devise suitable queries to show association between cytokine, SNPs and diseases.
- To disseminate the database over the web with suitable web interface so that the cytokine, SNP and disease data can be retrieved easily.
- To build a prediction tool to find disease association potential for newly discovered SNPs.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Hardware

All the work was done on a personal Windows Vista and Windows 7 with a 2 GHz Intel core 2 duo processor and 2 GB and 4GB of RAM respectively.

Software & Databases

The database contains data retrieved from different databases available online and from Pubmed literature through manual annotation. For the data classification supervised learning method tools were utilized for the classifiction purpose. Table 1 shows a list of databases and softwares used for the database creation and disease and non-disease classification of SNP based on function class.

Table 1.) List of software tools and databases used

atabase/Software	eb location/Programming language
MMPORT	ww.immport.org
BSNP	ww.ncbi.nlm.nih.gov/dbSNP
UBMED	ww.ncbi.nlm.nih.gov/pubmed
VM	BSVM
ERCEPTRON,MLP	ython code
DF	ython code
bot	CBI Eutils tool generates perl code; retrieves data from NCBI

3.2 Procedure

3.2.1 Data collection

Cytokine data collection

Cytokine data was collected from ImmPort (www.immport.org) an immunological portal containing information on immunological genes, pathways, networks, SNP information and genes and pathways associated with B cell and T cells and innate and adaptive immunity. The database currently has information on 456 cytokine genes. The cytokine data was retrieved in Excel and its content analyzed for classification into different cytokine classes before being added to the database. All the attributes of the data retrieved from ImmPort were maintained as it was appropriate annotation of the cytokine data that we wanted to add to the cytokine table of the database.

SNP data collection

SNP data was retrieved from the dbSNP database with the E-utils tools E-bot. A perl code generated from E-utils retrieved the dbSNP data in an XML file. This XML file was parsed with perl code in an appendix. The SNP data generated was further curated with Perl script to add rs identifier to the SNPId which is the unique identifier for the SNPs. The data was curated to add the chromosomal location for the SNP with missing chromosomal locations and function class was annotated to include functional and non-functional class annotation. Heterozygosity value and Heterozygosity standard error were curated to add the values as 0 for their missing values. The data was curated to remove the redundant values of SNPs and was cleansed of any repeats and missing values by using the assistance of excel methods.

Disease data collection

Disease data was collected from the PubMed literature cited in the dbSNP database. The data was manually retrieved by examining and reading each abstract to look for cytokine SNP association for disease and retrieving the relevant disease information from the PubMed abstract. The data was annotated and hyperlinked to the PubMed abstract. The data was cleansed to remove the duplicate values or missing values from the data in Excel by application of Excel functions.

Cleansing of the data

The data was cleansed with Microsoft Excel methods and functions. The data was cleaned to remove the duplicate rows and missing values. The attributes with wrong data was manually curated and missing data was added manually. For part of the additions and deletions, simple perl codes were applied to get data in proper format and correct values. All the attributes of the SNP, Cytokine and Disease table were manually examined to find the incorrect, repeated or missing values. They were either modified, corrected, deleted to ensure the consistency and accuracy of the data.

3.2.2 Annotation and Curation of the data

Data was annotated before being added to the database. Cytokine data retrieved from ImmPort was different for different cytokine groups and they were combined to make it one group. All the annotation was performed in Excel and the application of Venn diagrams present in the Array track tool [http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm]. The errors present in the data were curated by checking for repeated cytokines and any repetition of cytokines in the aliases. Cytokine data was properly annotated and curated but not much work was needed to be done as it was retrieved from annotated and curated ImmPort database. SNP and disease data required much annotation and curation. For the annotation of the SNP data the data was retrieved with an SQL query from the IU Knowledgebase dbSNP server (http://discern.uits.iu.edu:8421) more specifically http://discern.uits.iu.edu:8421/dbSNP/dbSNP/).   The SNP data was not very comprehensive as there were ~11,500 SNP data retrieved and comparing the gene list of SNP data to cytokine genes with Venn diagram suggested there were many cytokine genes for which SNP data was missing. This data was curated as it was retrieved by writing the SQL query such that only selected attributes were retrieved from the IU knowledgebase dbSNP JDBC server. The retrieved data was manually examined for correction of any errors present such as repeated data, missing data or incorrect gene data. Any incorrect information was either manually removed or curated by retrieving the correct value from dbSNP for the particular gene. SNP data was not complete and not comprehensive; hence it needed to be retrieved using other methods. Batch query search available online from dbSNP did not return the SNP data in the correct format and it did not return the desired attributes of SNP to be correctly added to the database as it would have required a voluminous amount of curation and annotation with Perl script. Hence, to ease up the process, data was retrieved with NCBI E-bot tool (http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi).The following steps were employed for data retrieval, annotation and curation.

i.) The list of cytokine genes was selected from the cytokine data retrieved from ImmPort database.

ii.) The cytokine list was submitted to the bioDBnet (http://biodbnet.abcc.ncifcrf.gov/) to change the cytokine list to dbSNP Id.

iii.) The dbSNP Id retrieved was changed from rsId to numeral identifier, SNPId list with a simple Perl script.

iv.) The retrieved SNPId list was submitted to the E-bot tool and perl code was generated to retrieve the full XML record. (Perl code can be located in appendix v)

v.) The retrieved record was parsed with a Perl code (SNPdataparsing.pl) to retrieve the curated SNP data with required SNP attributes.

vi.) The genes of the SNP data was analyzed with Array-track Venn-diagram tool (http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm) comparing it to ImmPort cytokine list to find the missing cytokine gene SNP data. Steps i-vi were repeated until the entire cytokine gene SNP data was retrieved.

vii.) The data was analyzed in Excel and part of it with perl script to remove the redundancy and locate the missing values. The data was manually examined to add the correct values for the incorrect data or to add the missing data from the dbSNP database. Most of the annotation and curation of the SNP attributes were done by the Perl code in step v. Remaining missing values, redundant values retrieved by the E-bot generated Perl code and repeated values generated by step vi data retrieval was manually curated and annotated. Each row and column of SNP data was examined manually and the correct annotation was added if missing by comparing it to the dbSNP database. Functional class annotation of almost all the cytokine SNPs was correctly retrieved and mostly there was error in SNP functional class data being retrieved in wrong columns and intermingled with other columns. This needed to be manually curated and corrected. Some function class annotation had extraneous information such as amino acid mutation information for missense function class, which needed to be removed. Also, there was reference function class SNP data with amino acid change information, and these SNPs were removed from the data. After addition of data to the database, the data was again examined for inconsistencies. And, with the phpMyAdmin tool the addition and correction of the data could be performed taking dbSNP as reference. Currently curated and annotated number of SNP data in the database is 63,356.

Disease data was retrieved from a PubMed literature link present for the SNP data in the dbSNP database. The entire cytokine gene SNPs with associated PubMed literature was examined and then the abstracts were manually read to retrieve the information to be included in the database. The disease and SNP association information, if suggested to be found with low p-values, were selected and the corresponding attribute information from the abstract was selected and added to the disease data in

Excel. This disease data was completely manually curated and annotated and was added to the database after careful examination of the data for repetition, missing or incorrect values. Once added to the database the data was again examined for inconsistencies as was done with SNP data and the corrected mistakes were eradicated. Later a web link record in PubMed database was added by adding the web-address of the PubMed Id to the Disease table and then by modifying the php script to provide the web link to the PMID. The disease data associated to cytokine gene SNP was completely manually curated and the attributes link the SNP attributes to the diseases, Cytokine genes, SNPId were verified by the web- linked PMID associated abstract. This annotation of disease data provides information on linkage of disease to cytokine gene and SNPs and vice versa.

Later on Cytokine, SNP and Disease tables and corresponding PHP pages were modified to add the NCBI cytokine gene address, SNPId dbSNP web address connecting to corresponding NCBI gene and SNP page.

3.2.3 Database Schema design

3.2.3.1 Data dictionary

The database dictionary of the database is described in Table 2.

Table 2.) Data description of the different attributes present in database

DACS-DB TABLE	ATTRIBUTE	DATATYPE	DATAVALUE
CYTOKINE	GeneId	VARCHAR	50
CYTOKINE	GeneSymbol	VARCHAR	50
CYTOKINE	Address	VARCHAR	200
CYTOKINE	CytokineFullName	VARCHAR	100
CYTOKINE	Aliases	VARCHAR	200
CYTOKINE	Chromosome	VARCHAR	20
SNP	Chromosome	VARCHAR	10
SNP	GeneSymbol	VARCHAR	50
SNP	GeneId	VARCHAR	10
SNP	SNPId	VARCHAR	20
SNP	ChromosomalLocation	VARCHAR	20
SNP	FunctionClass	VARCHAR	50
SNP	Allele	VARCHAR	10
SNP	HeterozygosityValue	VARCHAR	50
SNP	HetStdErr	VARCHAR	50
SNP	5primeSequence	blob	
SNP	3primeSequence	blob	
Disease	GeneId	VARCHAR	10
Disease	GeneSymbol	VARCHAR	20
Disease	Disease	VARCHAR	200
Disease	SNPId	VARCHAR	20

Disease	link	VARCHAR	100
Disease	Year	VARCHAR	4
Disease	PMID	VARCHAR	12

3.2.3.2 E-R diagram

An Entity-relationship diagram was generated to show the relationship between the Cytokine, SNP and Disease tables. It shows the relationships between the tables and the unique identity field which can connect the tables. These can be defined as the relationship establishing attributes which suggest whether one instance of an attribute in a table is connected with how many instances of the same attribute in a connecting table and vice-versa. There is one to one, one to many and many to many relationships between the tables. Figure 1 shows the entity relationship diagram for DACS-DB.

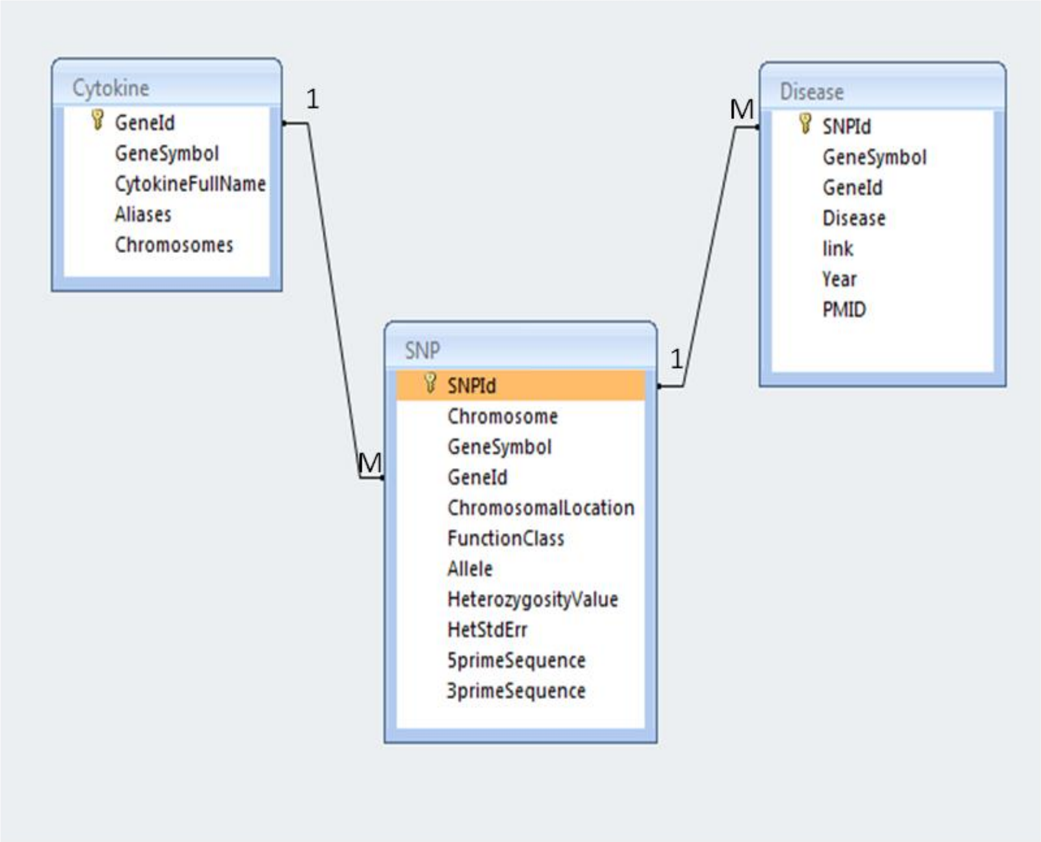


Fig. 1: Entity-relationship diagram: The relationship between Cytokine, SNP and Disease table is represented in the diagram. Cytokine table is connected to the SNP table with one to many relationships. The SNP table is connected to the disease table with one to many relationships.

3.2.4 Database creation

Creation of the database

The database was created on the MySQL platform. The database is built on the MySQL command prompt.

Creation of the tables

The database tables were built on the phpMyAdmin platform. The data dictionary describes the data type and data size of the attributes of the three tables.

3.2.5 Storage of the data

All the table values were added to the database from command line with INSERT code for adding values from text file.

3.2.6 Data definition language and Data manipulation language with PhpMyAdmin

Database tables were created and updated with DDL commands while data was added and deleted with application of DML commands.

3.2.7 Database server

The database is maintained by the Indiana University, Bloomington cluster of databases. It is maintained and updated by the university group of super computers and is the source of the data store house for commercial, academic and course work purposes.

Previous server

The database was previously maintained at https://libra45.uits.iu.edu port no. 3234, now it has been stopped.

Latest server

The database is currently maintained at https://rdc04.uits.iu.edu port no. 3234 and is the current active server and the database is maintained on this server.

Database server for DACS-DB

Current status of the server

The database is currently active and is only once a month not available online for monthly maintainanace.

3.2.8 Web end of the database

The database front end is built with PHP and is maintained by the webserve server of Indiana University.

3.2.9 Database Web end

The database webend is accessible at webserve.iu.edu and accessible at the url http://www.iupui.edu/~cytosnp.

Connecting the database to the web pages

The database webfront is connected to the database with PHP mysql_connect function and the interface for the webpage design purpose is provided by the WinSCP tool. The PHP codes written in text editor are saved in the web directory of the webserve and are accessible through the hypertext transfer protocol.

Different web pages present in the database

The database has the home page consisting of a header file, a brief note about the database, information of the visitor count and the description of the database author and university url. The database has webpages on Cytokines, SNP, Diseases, Submission Form and Help/FAQ.

Queries present for the web pages

The Cytokine page contains information on cytokine and search options with wild card character or cytokine search option word by word. The SNP page has search option for the search with Chromosome, GeneId, GeneSymbol, SNPId, and FunctionClass. The Disease page has search option available with GeneId, GeneSymbol, Disease, SNPId, Year, and PMID. The

Submission form web page provides information for submitting SNP, disease, and publication information submitted through the email Id. The Help/FAQ webpage contains information about the database and provides information about different tables and data and how it can be accessed with suitable queries.

3.2.10 Accessibility of the database

The database is accessible at the url http://www.iupui.edu/~cytosnp

3.3 Classification of the data by machine learning approach

3.3.1 Type and description of data analyzed

There are approximately 60,438 SNPs associated with 456 cytokines, and approximately 4,667 diseases associated with cytokine gene SNPs as cited in numerous PubMed articles. This work sought to fill the knowledge gap between disease-associated cytokine SNPs and disease-causing SNPs. The data was stored as searchable, curated, and annotated cytokine gene SNP data in the DACS-DB database. DACS-DB contains a catalog of manually curated and annotated cytokine genes, their SNPs, and the diseases associated with the corresponding cytokine SNPs. It has comprehensive information on important SNP attributes such as function class and heterozygosity value which can be important in understanding the causal association of a given SNP to disease. The database is publicly available as a web resource at http://www.iupui.edu/~cytosnp/.

3.3.1 Data for SVM

The data was balanced (to ensure the number of records in each class as the same) before taking it into any tool. In doing so, we ended up with 864 records of each class (disease/non-disease). However, a further requirement of SAS Multiple Logistic Regression imposed was in regard to cell sizes based on categorical variables such as our Function Class term. Using the two categories (coding/non-coding) for the Function Class, the cell break down was:

- Non-coding: non-disease (851), disease(799)
- Coding: non-disease(13), disease(65)

In order for SAS to accurately create a model, it must have a minimum of 10 observations per feature (e.g. 10 observations X 2 features = 20) in the smallest cell. Therefore, our smallest cell of 13 was not sufficient for two features. With this knowledge, we set out to determine why we had so few Coding/Non-disease records. Another analysis of the data revealed the reason:

- We have many more non-disease records (44,724) than disease records (864)
- The vast majority of our non-disease records are non-coding (44,166) versus coding (558)
- When we take a random sample of 864 non-disease records (to match the number of disease records), we then naturally only take a few coding records

3.3.2 Retrieval of the data

The data for the classification purpose was retrieved from DACS-DB database by the application of the following SQL query.

Query used to extract data from the DACS-DB

```
SELECT SNP1.SNPId, SNP1.FunctionClass, SNP1.HeterozygosityValue, Disease.Disease

FROM SNP1 LEFT OUTER JOIN Disease ON SNP1.SNPId = Disease.SNPId;
```

3.3.3 Cleansing of the data

Table 3 shows steps that were applied for the cleansing and classification purpose.

Table 3: Steps involved in SNP data cleaning, analysis and classification.

Task
1. Develop and execute a query against the DACS-DB SNP and Disease tables to capture SNP ID, Function Class, Heterozygosity Value, and Disease. Place extracted data in an Excel spreadsheet (See Appendix 1 for SQL query).
2. Write a Python script to pre-process the extracted data to:
a. Remove unwanted rows
b. Translate Function Code values from text to numeric
c. Translate the values of Disease from text to numeric, with class 0 = not known to be associated with a disease, and class 1 = known to be associated with a disease
d. Randomly select an equal (to the number of rows in the smaller class) number of rows from the larger class and combine with smaller class to create balanced data
e. Randomize the balanced data
f. Plot the features to determine if they are linearly separable
g. Separate the data into 50% training, 25% validation, and 25% testing
3. If linearly separable, train, validate, and test data with Perceptron
4. If not linearly separable, train, validate, and test data with Multi-Layer Perceptron
5. Use a confusion matrix to indicate correct classification percentage
6. Time permitting, download Support Vector Machine software from the web, train and test the data with SVM

3.3.4 Classification methods: A Different Approach

Data classification begun by analyzing whether there was any good relationship between Function Class/Heterozygosity Value and association with a disease. Statistical, Multiple Logistic Regressions (MLR), RBF, LDA, SVM were applied for the purpose. There was no linear relationship between dependent (disease/no-disease) and multiple independent values. This type of problem applies when there is a binary dependent value (disease/no-disease) and multiple independent values (Function Class, Heterozygosity Value). The purpose of the approach was to identify a good MLR model, then translate that back (add non-linear terms) to our machine learning model.

Data input was prepared for a very sophisticated commercial statistical software package, SAS, Version 9.2. Different variations of different Function Classes that we had used with the MLP (weighted values, coding/non-coding values, and unique values) along with Heterozygosity Value were applied. For all of the combinations, SAS indicated that the data did not fit the model. These findings confirmed we were giving SAS a linear model, but the data did not support that model. In other words, there was no linear relationship between Function Class/Heterozygosity Value and Disease.

Then data was normalized to improve the linearity of the data. The first method was transforming the Heterozygosity Value by taking its natural log. Second, was adding interaction terms. In this case, a value was added to take into SAS that is the product of two existing values such as Function Class X Heterozygosity Value; several combinations of values were tried. In all cases, the data still did not support the model. The final technique used with SAS was where terms were squared and cubed (e.g. Heterozygosity Value squared and Heterozygosity Value cubed). Once again, SAS indicated the data still did not support the model. The new approach had only served to verify the relationship between the data and was not linear.

Examining the Data

Although statistical software did not provide any relationship between the input terms, it was noticed we were keeping a record for each disease to which the SNP was associated. Therefore, the data was over-representing the fact that a given SNP was associated with a disease. To care for this mistake, logic to include a given SNP only once, no matter how many diseases with which it is associated was added. Next, the MLP was run again, but this time with the new, smaller dataset. We first ran with the weighted Function Classes. The error values were about half of what they were before using the smaller dataset. Amazingly, the best MLP was now with 5 hidden units. However, the Confusion Matrix was not very good with only 66% correct classifications. We then ran the MLP with the coding/non-coding Function Classes. Once again, the best number of MLP hidden nodes was much lower. However, the Confusion Matrix was still not that good with only 68% correct classifications. Finally, we ran the MLP with the unique Function Classes. After several iterations, the best number of MLP hidden nodes turned out to be 22. The Confusion Matrix was slightly worse at just less than 67% correct classifications. Based on these results, the coding/non-coding Function Classes, while not good, were the best classifiers. Just as an experiment, we used this model with the log of the Heterozygosity Value rather than just the value. The percent of correct classifications was worse at 59%.

We ran RBF (with the coding/non-coding Function Classes) several times using the parameters from the Stephen Marshland, Machine learning book;. The best percent correct of classifications was 68%, the same as the MLP. Next, we ran the Linear Discrimant Analysis from the Stephen Marshland, Machine learning book with the coding/non-coding Function Classes; It only produced a plot, not a confusion matrix. Therefore, it was to be determined how to use its output. The way LDA works, it simply modifies the data, not the targets. Consequently, we took its modified data into the same MLP code as used above. We found the best number of hidden nodes to be 22; then MLP was run a few times. The percent of correct classification was comparable to just the MLP using non-LDA input at about 68%.

3.3.5 Support Vector Machine

After analyzing data with the supervised machine learning tools, data was analyzed with the Support Vector Machine (SVM) tool, LIBSVM Release 3.1. (it included Windows, Python, and Java solutions among many others. It includes four kernels: linear, polynomial, radial basis function, and sigmoid. RBF kernel was chosen for analyzing data based on the recommendation of the LIBSVM guide. LIBSVM also required another software program for graphing to be installed called gnu plot.

After installing the software, data was added in the format required by LIBSVM. Its format was considerably different, requiring a nearly new Python script. Completing this step for the coding/non-coding version of the data, checkdata.py script was run. Both our training and testing datasets passed its checks. Also included with LIBSVM was a Python script called easy.py that automates the entire process of scaling the data, selecting parameters, choosing a model, training the model, and testing the model. The documentation indicated it works well for moderately complex data. However, it did not work that well for the data. It took about half an hour, on the second run; the correct classification percent returned was just under 65% and the second run's percent was nearly the same. Nevertheless, there was still hope for improvement. The guide listed a method of improving this score by running individual processes separately, optimizing each of them. However, even before running LIBSVM manually, additional changes were made to data (per the LIBSVM guide) and received a score of almost 73%, using easy.py. The data was normalized in Python data preparation script before taking it into the easy.py script and the data was scaled in the training dataset and the testing dataset via the LIBSVM Python script, scale.py. The scaled dataset was taken into easy.py. The result was the best correct classification percentage, 72.9%. Also training LIBSVM using combined dataset (both training and testing datasets) as input to the LIBSVM grid.py script, the result was an even higher 73.8%.

CHAPTER FOUR

RESULTS

4.1 Database result

DACS-DB database is available online and can be accessed at the web address http://www.iupui.edu/~cytosnp . The database is available all the time barring once monthly maintenance of the database. It can be a source or repository for immunologists, geneticists and biomedical researchers to understand the role of genetics behind the cause of diseases and characterize the drug targets for the diseases.

Query option result and submission available

Database has query option available for Cytokine, SNP and Disease tables as shown in the table 4, while the submission page has the option of submitting disease, gene , SNP with literature citation to the database through email.

Table 4.) List of different attributes present in the database with query option to the database.

Table	Query
Cytokine	GeneId
Cytokine	GeneSymbol
Cytokine	CytokineFullName
Cytokine	Aliases
Cytokine	Chromosome
SNP	Chromosome
SNP	GeneSymbol
SNP	GeneId
SNP	SNPId
SNP	FunctionClass
Diseases	GeneId
Diseases	GeneSymbol
Diseases	Disease
Diseases	SNPId
Diseases	Year
Diseases	PMID

4.2 Cytokine table result

Search options

Cytokine table has search option available for GeneId, GeneSymbol, CytokineFullName, Aliases, Chromosome. Cytokine table also has a search option available with wild card search. GeneId, GeneSymbol, Aliases can be searched with partial word search.

Search results

The result of the search is the number of the rows of the data from the table for which the search term matches the query in the database. The search result displays the result of the search terms. All the 5 attributes of the search term are displayed in the search result along with information about the corresponding attributes in the result page.

Cytokine statistics

The Cytokine page contains information on 456 Cytokine genes. The following is the statistics and pie chart representing the different type of cytokine information and the percentage distribution of the Cytokines.

Table 5.) List of classes of cytokines and their number and percentage distribution

Cytokine	Total no.	Percentage
Interleukins	47	10%
Interferons	17	4%
TNF-family	12	3%
TGF-β family	33	7%
Chemokines	102	22%
Growth hormones	245	45%

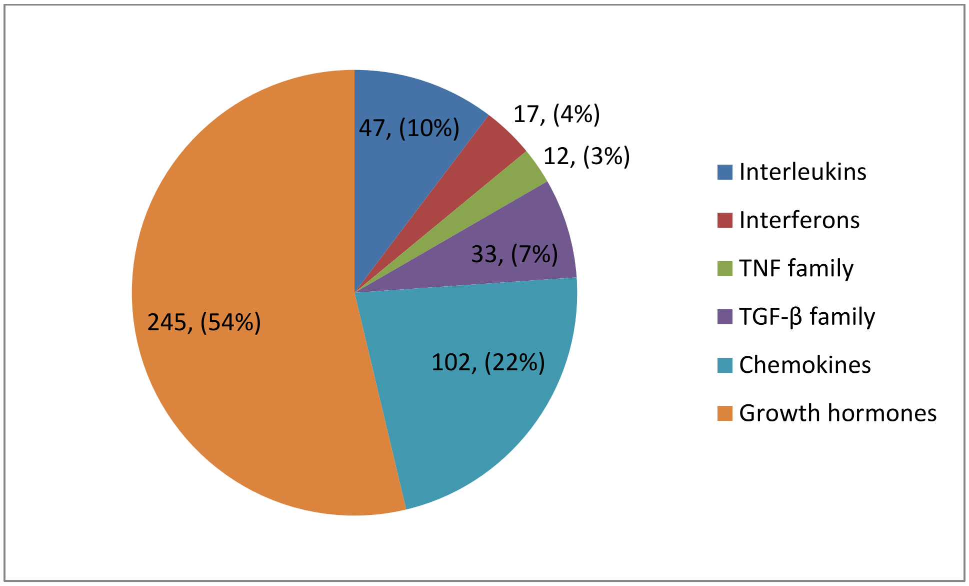


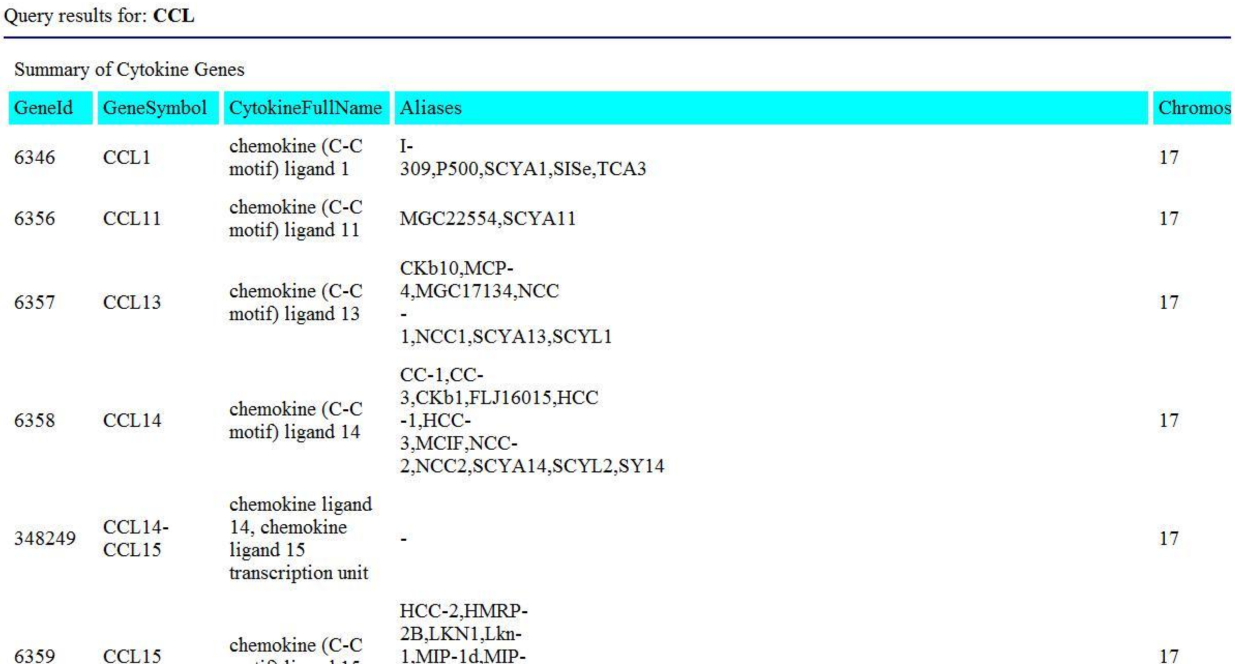
Fig. 2) *Pie chart distribution*, Pie-chart of classes of cytokines and number and percentage distribution of different cytokines present

Cytokine attributes result

Cytokines contain information about GeneId, GeneSymbol, CytokineFullName, Aliases, Chromosomes. These attributes provide information about the annotation of the gene and different aliases present for the gene to search for the Cytokine.Each cytokine gene symbol is connected to a NCBI gene page which can provide information about the Cytokine genes.

Cytokine Search result

The following two results display the search result of Chemokines search with CCL as search keyword and IL8 gene from the two search options present in the Cytokine page.

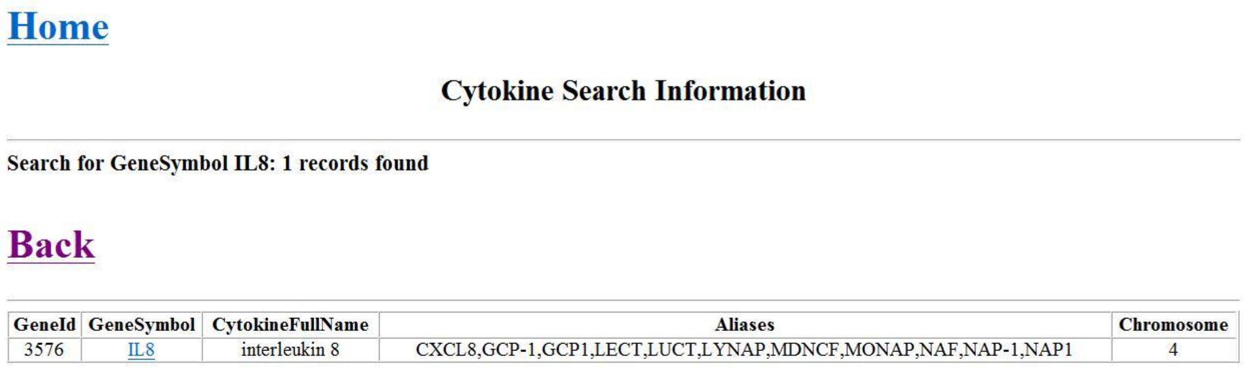


Query results for: **CCL**

Summary of Cytokine Genes

GeneId	GeneSymbol	CytokineFullName	Aliases	Chromos
6346	CCL1	chemokine (C-C motif) ligand 1	I-309,P500,SCYA1,SISe,TCA3	17
6356	CCL11	chemokine (C-C motif) ligand 11	MGC22554,SCYA11	17
6357	CCL13	chemokine (C-C motif) ligand 13	CKb10,MCP-4,MGC17134,NCC-1,NCC1,SCYA13,SCYL1	17
6358	CCL14	chemokine (C-C motif) ligand 14	CC-1,CC-3,CKb1,FLJ16015,HCC-1,HCC-3,MCIF,NCC-2,NCC2,SCYA14,SCYL2,SY14	17
348249	CCL14-CCL15	chemokine ligand 14, chemokine ligand 15 transcription unit	-	17
6359	CCL15	chemokine (C-C	HCC-2,HMRP-2B,LKN1,Lkn-1,MIP-1d,MIP-	17

Fig. 3: *Cytokine Search result*, Search result of CCL search in the Cytokine search page showing different attribute results returned with the wild card search option mainly done to boost the search of cytokines with aliases.



Home

Cytokine Search Information

Search for GeneSymbol IL8: 1 records found

Back

GeneId	GeneSymbol	CytokineFullName	Aliases	Chromosome
3576	IL8	interleukin 8	CXCL8,GCP-1,GCP1,LECT,LUCT,LYNAP,MDNCF,MONAP,NAF,NAP-1,NAP1	4

Fig. 4, *Cytokine Search result*: Search result of IL8 gene search in the cytokine table showing result of IL8 gene search in the cytokine table.

4.3 SNP TABLE RESULT

Search options

SNP table has search option present for GeneId, GeneSymbol, Chromosome and FunctionClass. SNP data can be searched for any of these attributes and it returns the result for the number of rows present for the search term matching the search attributes value to the data in the database.

Search results

SNP search returns the result for the search term matching the result for the query term matching the term in the database and returns the information of the 12 SNP attribute information present in the database. The result provides information on the annotation result of the SNP data present for the number of the SNP data returned for the query term. This provide important information about the

SNPs attribute characteristics such as Function class, Heterozygosity value, Heterozygosity standard error, Allele, Chromosomal position and sequence information.

SNP statistics

The SNP page contains information of the 456 Cytokine genes. The following is the statistics and pie chart representing the SNP data of different types of cytokine genes and the percentage distribution of the SNPs of the cytokines.

Table 6.) List of SNP data present for Cytokine classes and percentage distribution

Cytokines	No. of SNPs	% of SNPs
Interleukins	4026	7
Interferons	679	1
TNF-family	1019	2
TGF-β family	4426	7
Chemokines	13372	22
Growth hormones	36944	61

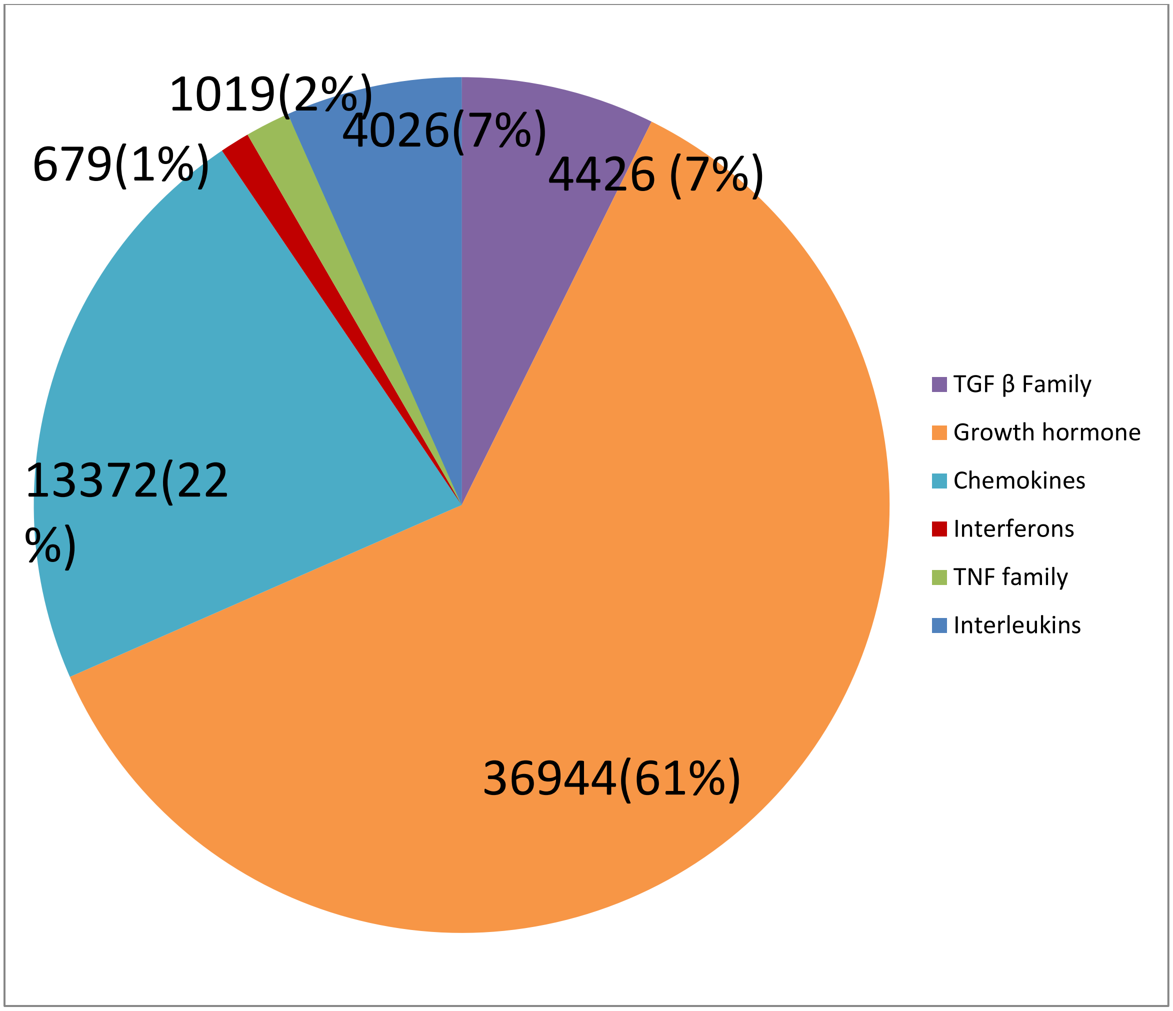


Fig. 5 *Pie chart distribution*, distribution of SNP data present in different cytokine classes and percentage distribution of SNPs data present in different cytokine classes.

SNP statistics with information on average number of SNP per disease with publication and average number of SNP present without publication (Table 7): There is an average of 71 SNPs present for all diseases in the database (top row) while among the diseases with literature evidence for association to a SNP (PubMed), this average reduces to two (bottom row). This latter average will increase as more publications implying disease-SNP associations are published.

Table 7.) Average number of SNPs present in DACS-DB diseases associated with SNP data present in SNP table and SNP data with publication in Disease table

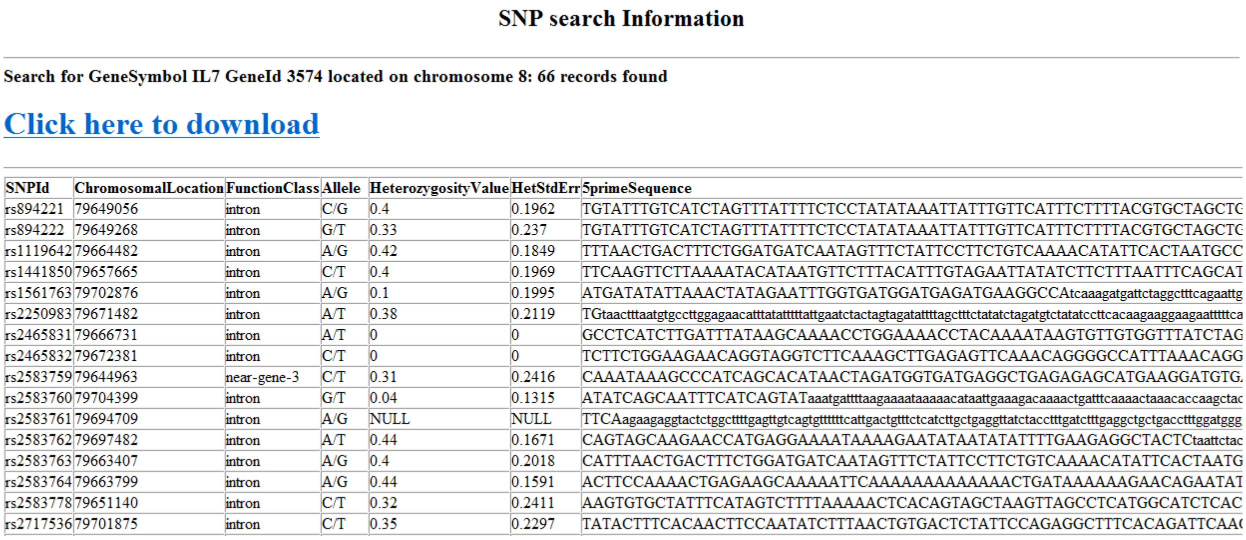
Total no. of distinct diseases present in Disease table	Total no. of SNP present in SNP table and Disease table	Average SNP per disease present in SNP and Disease table
892	63356	71
892	1734	2

SNP attributes result and download option

SNP has attribute information on Chromosome, GeneSymbol, GeneId, SNPId, Chromosomal Location, Function Class, Allele, Heterozygosity Value, Heterozygosity standard error, 5 prime Sequence, 3 prime Sequence. The SNP data can be downloaded in the Excel.

SNP search result

The following search result displays the search result of searching IL7 gene as search query term.



SNP search Information

Search for GeneSymbol IL7 GeneId 3574 located on chromosome 8: 66 records found

Click here to download

SNPId	ChromosomalLocation	FunctionClass	Allele	HeterozygosityValue	HetStdErr	5primeSequence
rs894221	79649056	intron	C/G	0.4	0.1962	TGTATTTGTCATCTAGTTTATTTTCTCCTATATAAATTATTTGTTCATTTCTTTTACGTGCTAGCTG
rs894222	79649268	intron	G/T	0.33	0.237	TGTATTTGTCATCTAGTTTATTTTCTCCTATATAAATTATTTGTTCATTTCTTTTACGTGCTAGCTG
rs1119642	79664482	intron	A/G	0.42	0.1849	TTTAACTGACTTTCTGGATGATCAATAGTTTCTATTCCTTCTGTCAAAACATATTCACTAATGCC
rs1441850	79657665	intron	C/T	0.4	0.1969	TTCAAGTTCTTAAAATACATAATGTTCTTTACATTTGTAGAATTATATCTTCTTTAATTTCAGCAT
rs1561763	79702876	intron	A/G	0.1	0.1995	ATGATATATTAAACTATAGAATTTGGTGATGGATGAGATGAAGGCCAtcaaagatgattctaggctttcagaattg
rs2250983	79671482	intron	A/T	0.38	0.2119	TGtaactttaatgtgccttggagaacatttatatttttattgaatctactagtagatattttagctttctatatctagatgtctatatccttcacaagaaggaagaatttttca
rs2465831	79666731	intron	A/T	0	0	GCCTCATCTTGATTTATAAGCAAAACCTGGAAAACCTACAAAATAAGTGTTGTGGTTTATCTAG
rs2465832	79672381	intron	C/T	0	0	TCTTCTGGAAGAACAGGTAGGTCTTCAAAGCTTGAGAGTTCAAACAGGGGCCATTTAAACAGG
rs2583759	79644963	near-gene-3	C/T	0.31	0.2416	CAAATAAAGCCCATCAGCACATAACTAGATGGTGATGAGGCTGAGAGAGCATGAAGGATGTG
rs2583760	79704399	intron	G/T	0.04	0.1315	ATATCAGCAATTTCATCAGTATaaatgattttaagaaaataaaaacataattgaaagacaaaactgatttcaaaactaaacaccaagctac
rs2583761	79694709	intron	A/G	NULL	NULL	TTCAagaagaggtactctggcttttgagttgtcagtgtttttcattgactgtttctcatcttgctgaggttatctacctttgatctttgaggctgctgacctttggatggg
rs2583762	79697482	intron	A/T	0.44	0.1671	CAGTAGCAAGAACCATGAGGAAAATAAAAGAATATAATATATTTTGAAGAGGCTACTCtaattctac
rs2583763	79663407	intron	A/G	0.4	0.2018	CATTTAACTGACTTTCTGGATGATCAATAGTTTCTATTCCTTCTGTCAAAACATATTCACTAATG
rs2583764	79663799	intron	A/G	0.44	0.1591	ACTTCCAAAACTGAGAAGCAAAAATTCAAAAAAAAAAAAAACTGATAAAAAAGAACAGAATAT
rs2583778	79651140	intron	C/T	0.32	0.2411	AAGTGTGCTATTTCATAGTCTTTTAAAAACTCACAGTAGCTAAGTTAGCCTCATGGCATCTCAC
rs2717536	79701875	intron	C/T	0.35	0.2297	TATACTTTCACAACTTCCAATATCTTTAACTGTGACTCTATTCCAGAGGCTTTCACAGATTCAA

Fig.6 *SNP Search result*, Search result of IL7 gene in SNP table

4.4 DISEASE TABLE RESULT

Search options

Diseases have the search option available for GeneId, GeneSymbol, Disease, SNPId, Year, and PMID. Disease data can be searched for any of these attributes and it returns the result for the number of rows present for the search term matching the search attribute value matching to the attribute data in the database.

Search results

SNP search returns the result for the search term matching the result for the query term matching the term in the database and returns the information of the 6 disease attribute information present in the database. The result provides information on the annotation result of the disease for the number of the Disease data returned for the query term. This provides important information about the disease attributes as GeneId, GeneSymbol, Disease, SNPId, Year, and PMID. The Disease data for the search term GeneSymbol and Disease type has download option available in Excel.

Disease statistics

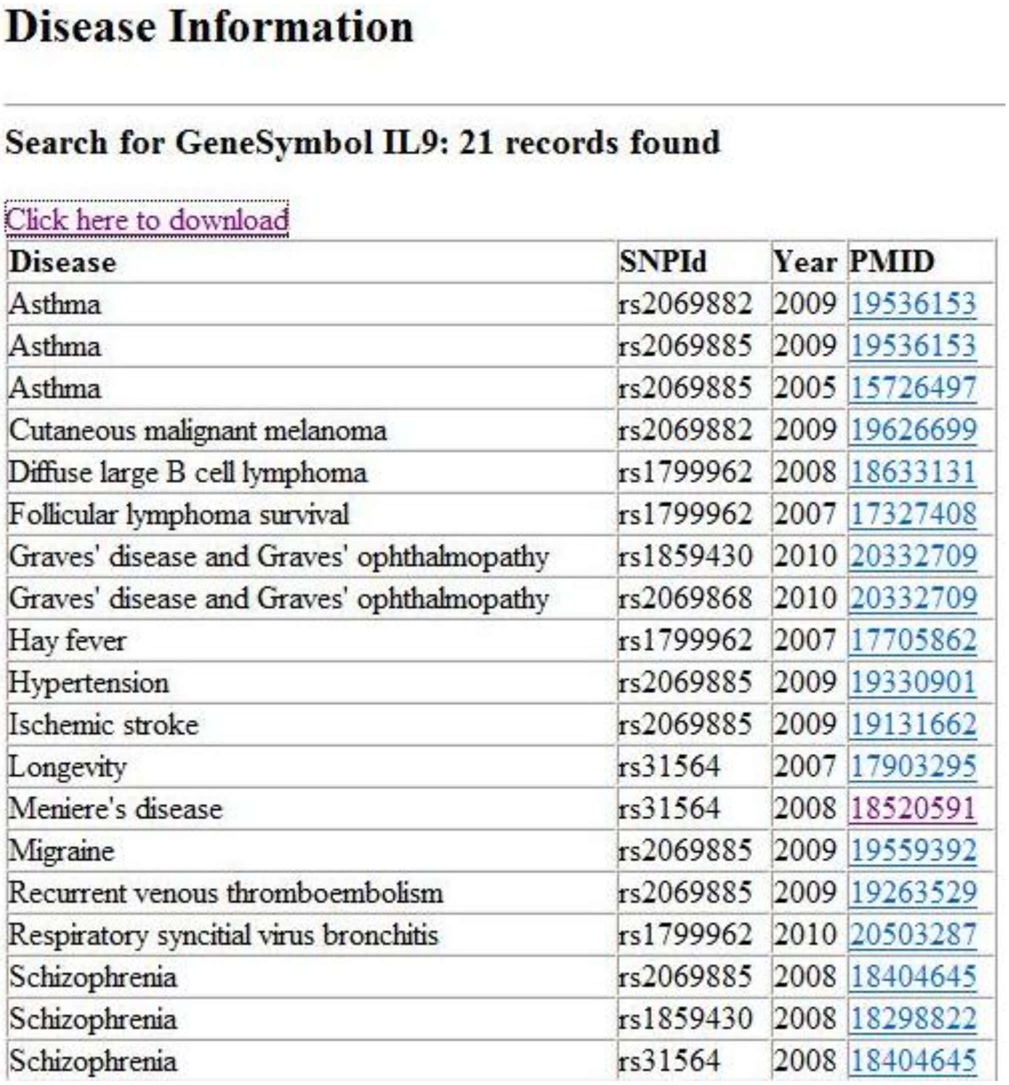
The disease table has information about 4669 PubMed literature data present for 456 cytokine genes. For all the disease data associated with disease there is a PubMed citation for the disease association with SNP. There is a total of 1,469 distinct PMID present for disease data and 1,737 distinct SNPId data associated with diseases present for the total SNP data. There is a total of 892 distinct disease data present for the 456 cytokine genes from the literature study.

Disease attributes result and downloads option

The database provides information about the Disease attributes as GeneId, GeneSymbol, Disease, SNPId, Year, and PMID. The Disease data for the search term GeneSymbol and Disease type has a download option available and it can be downloaded in Excel.

Disease search result

Figure 7 displays the search result of searching diseases associated with IL9 genes, while figure 8 shows the search result for the disease Multiple sclerosis and figure 9 the result for SNPId rs20451.Figure 10 displays the result of searching 5q31 region associated cytokine, SNPs and diseases with PubMed references.



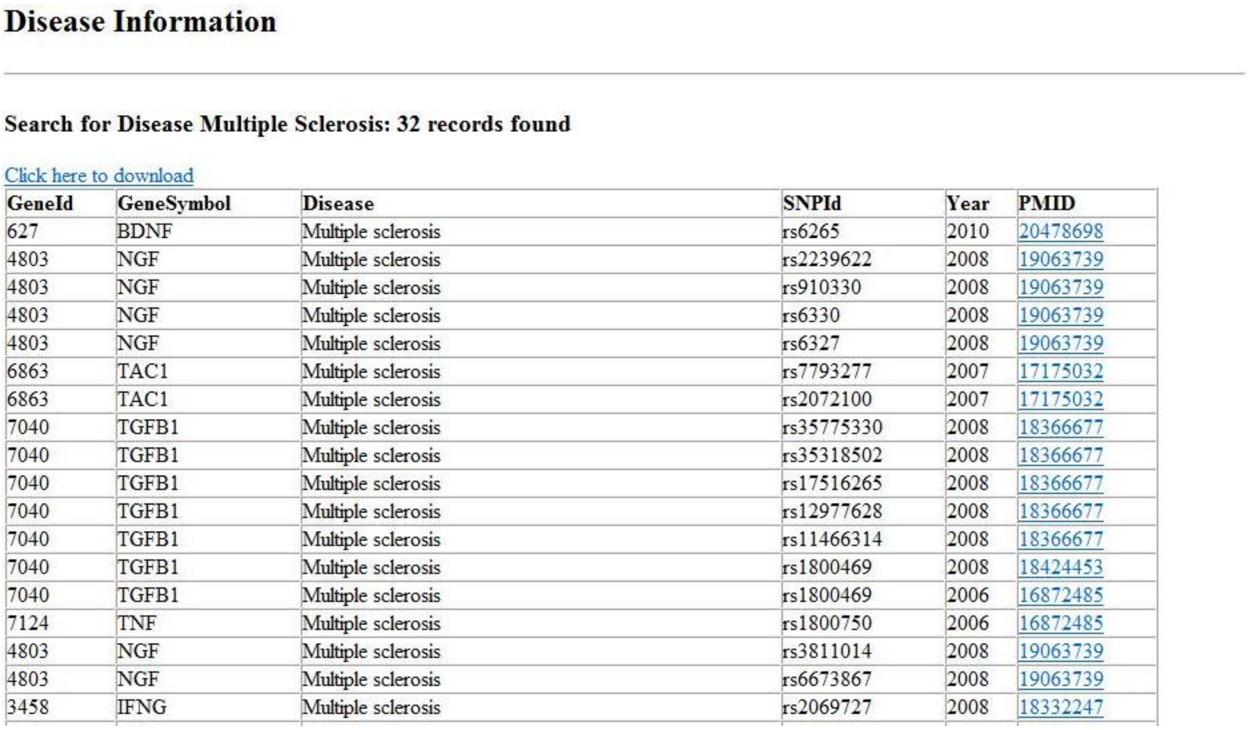
Disease Information

Search for GeneSymbol IL9: 21 records found

Click here to download

Disease	SNPId	Year	PMID
Asthma	rs2069882	2009	19536153
Asthma	rs2069885	2009	19536153
Asthma	rs2069885	2005	15726497
Cutaneous malignant melanoma	rs2069882	2009	19626699
Diffuse large B cell lymphoma	rs1799962	2008	18633131
Follicular lymphoma survival	rs1799962	2007	17327408
Graves' disease and Graves' ophthalmopathy	rs1859430	2010	20332709
Graves' disease and Graves' ophthalmopathy	rs2069868	2010	20332709
Hay fever	rs1799962	2007	17705862
Hypertension	rs2069885	2009	19330901
Ischemic stroke	rs2069885	2009	19131662
Longevity	rs31564	2007	17903295
Meniere's disease	rs31564	2008	18520591
Migraine	rs2069885	2009	19559392
Recurrent venous thromboembolism	rs2069885	2009	19263529
Respiratory syncitial virus bronchitis	rs1799962	2010	20503287
Schizophrenia	rs2069885	2008	18404645
Schizophrenia	rs1859430	2008	18298822
Schizophrenia	rs31564	2008	18404645

Fig. 7 *Disease search information*, Search result of searching IL7 gene in Disease table.



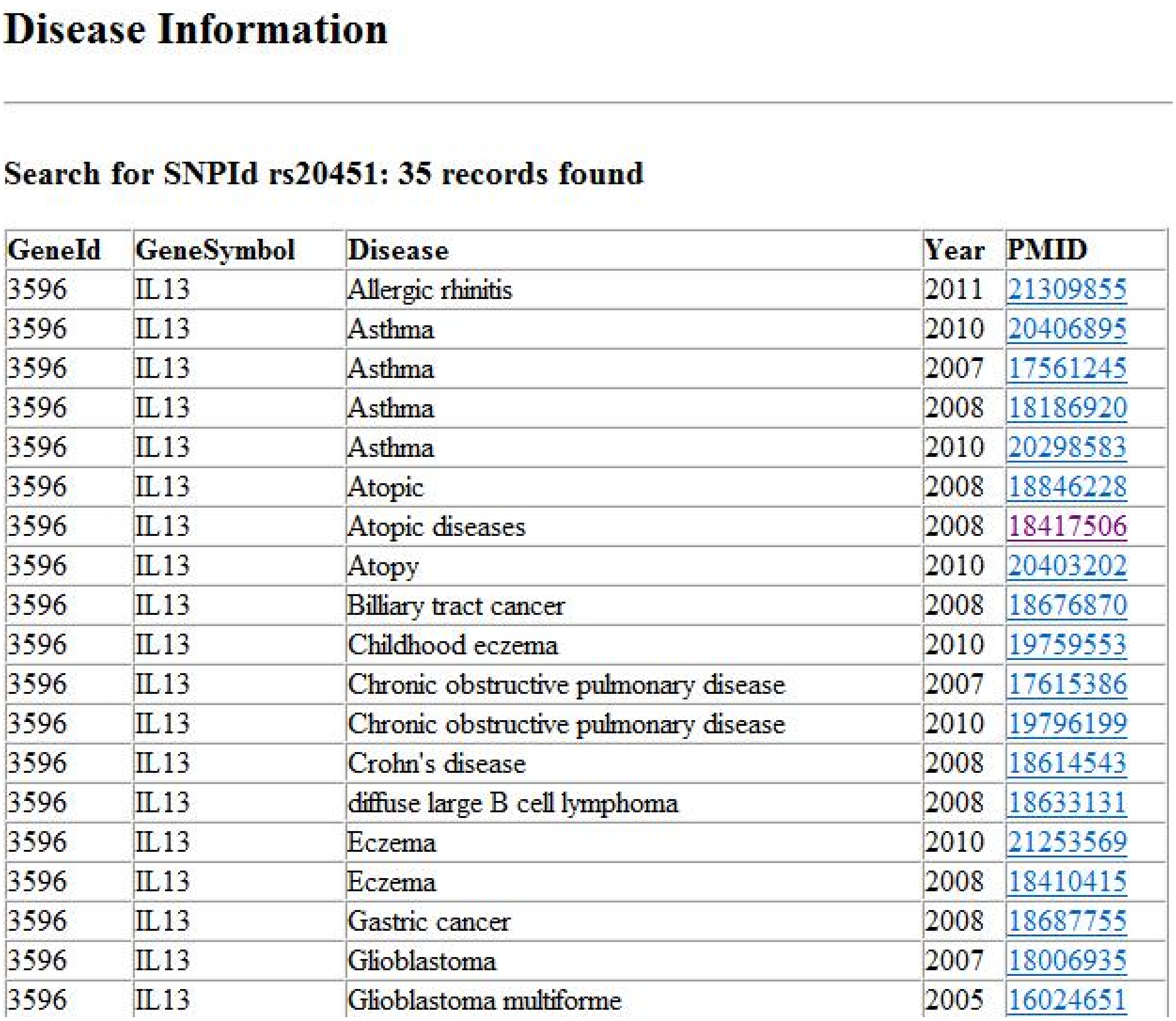
Disease Information

Search for Disease Multiple Sclerosis: 32 records found

Click here to download

GeneId	GeneSymbol	Disease	SNPId	Year	PMID
627	BDNF	Multiple sclerosis	rs6265	2010	20478698
4803	NGF	Multiple sclerosis	rs2239622	2008	19063739
4803	NGF	Multiple sclerosis	rs910330	2008	19063739
4803	NGF	Multiple sclerosis	rs6330	2008	19063739
4803	NGF	Multiple sclerosis	rs6327	2008	19063739
6863	TAC1	Multiple sclerosis	rs7793277	2007	17175032
6863	TAC1	Multiple sclerosis	rs2072100	2007	17175032
7040	TGFB1	Multiple sclerosis	rs35775330	2008	18366677
7040	TGFB1	Multiple sclerosis	rs35318502	2008	18366677
7040	TGFB1	Multiple sclerosis	rs17516265	2008	18366677
7040	TGFB1	Multiple sclerosis	rs12977628	2008	18366677
7040	TGFB1	Multiple sclerosis	rs11466314	2008	18366677
7040	TGFB1	Multiple sclerosis	rs1800469	2008	18424453
7040	TGFB1	Multiple sclerosis	rs1800469	2006	16872485
7124	TNF	Multiple sclerosis	rs1800750	2006	16872485
4803	NGF	Multiple sclerosis	rs3811014	2008	19063739
4803	NGF	Multiple sclerosis	rs6673867	2008	19063739
3458	IFNG	Multiple sclerosis	rs2069727	2008	18332247

Fig. 8 *Disease Search Information*, Search result of searching Multiple sclerosis disease in Disease table



Disease Information

Search for SNPId rs20451: 35 records found

GeneId	GeneSymbol	Disease	Year	PMID
3596	IL13	Allergic rhinitis	2011	21309855
3596	IL13	Asthma	2010	20406895
3596	IL13	Asthma	2007	17561245
3596	IL13	Asthma	2008	18186920
3596	IL13	Asthma	2010	20298583
3596	IL13	Atopic	2008	18846228
3596	IL13	Atopic diseases	2008	18417506
3596	IL13	Atopy	2010	20403202
3596	IL13	Billiary tract cancer	2008	18676870
3596	IL13	Childhood eczema	2010	19759553
3596	IL13	Chronic obstructive pulmonary disease	2007	17615386
3596	IL13	Chronic obstructive pulmonary disease	2010	19796199
3596	IL13	Crohn's disease	2008	18614543
3596	IL13	diffuse large B cell lymphoma	2008	18633131
3596	IL13	Eczema	2010	21253569
3596	IL13	Eczema	2008	18410415
3596	IL13	Gastric cancer	2008	18687755
3596	IL13	Glioblastoma	2007	18006935
3596	IL13	Glioblastoma multiforme	2005	16024651

Fig. 9, *SNP search information in Disease table*, Result of searching rs20451 SNP in the Disease table

Table 8, ***5q31 region disease association,*** Results of the cytokine genes and SNP associated diseases and references in 5q31 region are shown in the figure.

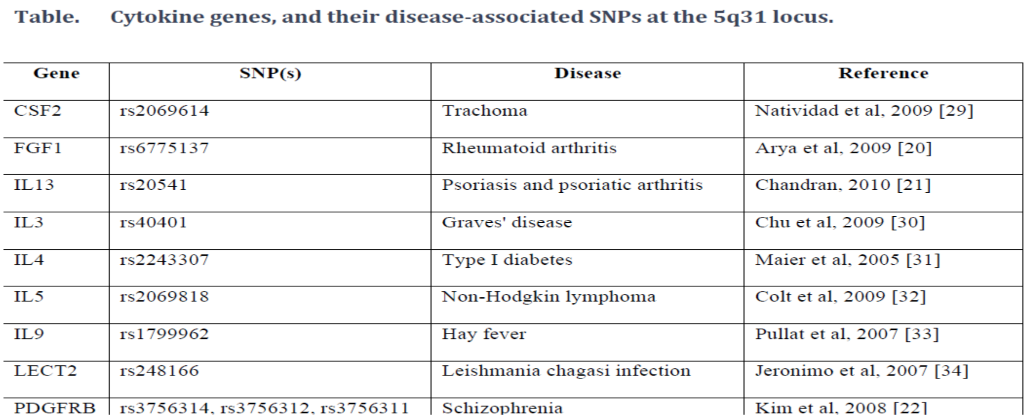


Table. Cytokine genes, and their disease-associated SNPs at the 5q31 locus.

Gene	SNP(s)	Disease	Reference
CSF2	rs2069614	Trachoma	Natividad et al, 2009 [29]
FGF1	rs6775137	Rheumatoid arthritis	Arya et al, 2009 [20]
IL13	rs20541	Psoriasis and psoriatic arthritis	Chandran, 2010 [21]
IL3	rs40401	Graves' disease	Chu et al, 2009 [30]
IL4	rs2243307	Type I diabetes	Maier et al, 2005 [31]
IL5	rs2069818	Non-Hodgkin lymphoma	Colt et al, 2009 [32]
IL9	rs1799962	Hay fever	Pullat et al, 2007 [33]
LECT2	rs248166	Leishmania chagasi infection	Jeronimo et al, 2007 [34]
PDGFRB	rs3756314, rs3756312, rs3756311	Schizophrenia	Kim et al, 2008 [22]

4.5 Classification result of SNP into disease and non-disease class

The results of different predictors and classification methods (supervised learning method) are shown below; the best overall correct classification percent was achieved by the SVM, using the coding/non-coding version of Function Class. However, both the MLP and LDA methods produced results that were nearly as good.

Python scripts and methods returned results which varied after multiple runs due to:

- The random selection of the class with less data to balance the number of records from each class (disease/non-disease)
- The random ordering and also selection of training, validation, and testing subsets

Table 9, Classification with different classification tools showing percent correct classification

				Percent	
Method/Tool	**Function Class Values**	**Normalization Method**	**Best Network Size**	**Correct**	**Best in Method**
MLP	coding-non-coding	unit variance	21	66.20	
MLP	coding-non-coding	maximum - minimum	50	71.30	X
MLP	weighted	unit variance	3	65.74	
MLP	weighted	maximum - minimum	4	66.20	
MLP	unique	unit variance	2	69.44	
MLP	unique	maximum - minimum	15	68.52	
RBF	coding-non-coding	unit variance	N/A	50.35	
RBF	coding-non-coding	maximum - minimum	N/A	65.05	X
RBF	weighted	unit variance	N/A	56.37	
RBF	weighted	maximum - minimum	N/A	59.38	
RBF	unique	unit variance	N/A	50.69	
RBF	unique	maximum - minimum	N/A	55.56	
LDA/MLP	coding-non-coding	unit variance	2	66.20	
LDA/MLP	coding-non-coding	maximum - minimum	3	66.90	
LDA/MLP	weighted	unit variance	50	72.69	X
LDA/MLP	weighted	maximum - minimum	3	69.44	
LDA/MLP	unique	unit variance	1	69.44	
LDA/MLP	unique	maximum - minimum	3	67.59	
SVM	coding-non-coding	SVM	N/A	73.84	X
SVM	weighted	SVM	N/A	66.67	
SVM	unique	SVM	N/A	64.81	

DATA NORMALIZATION RESULT

Figure 11 represents the classification of the data into disease and non-disease class taking the weighted value of function class into consideration. The data is not linearly separable.

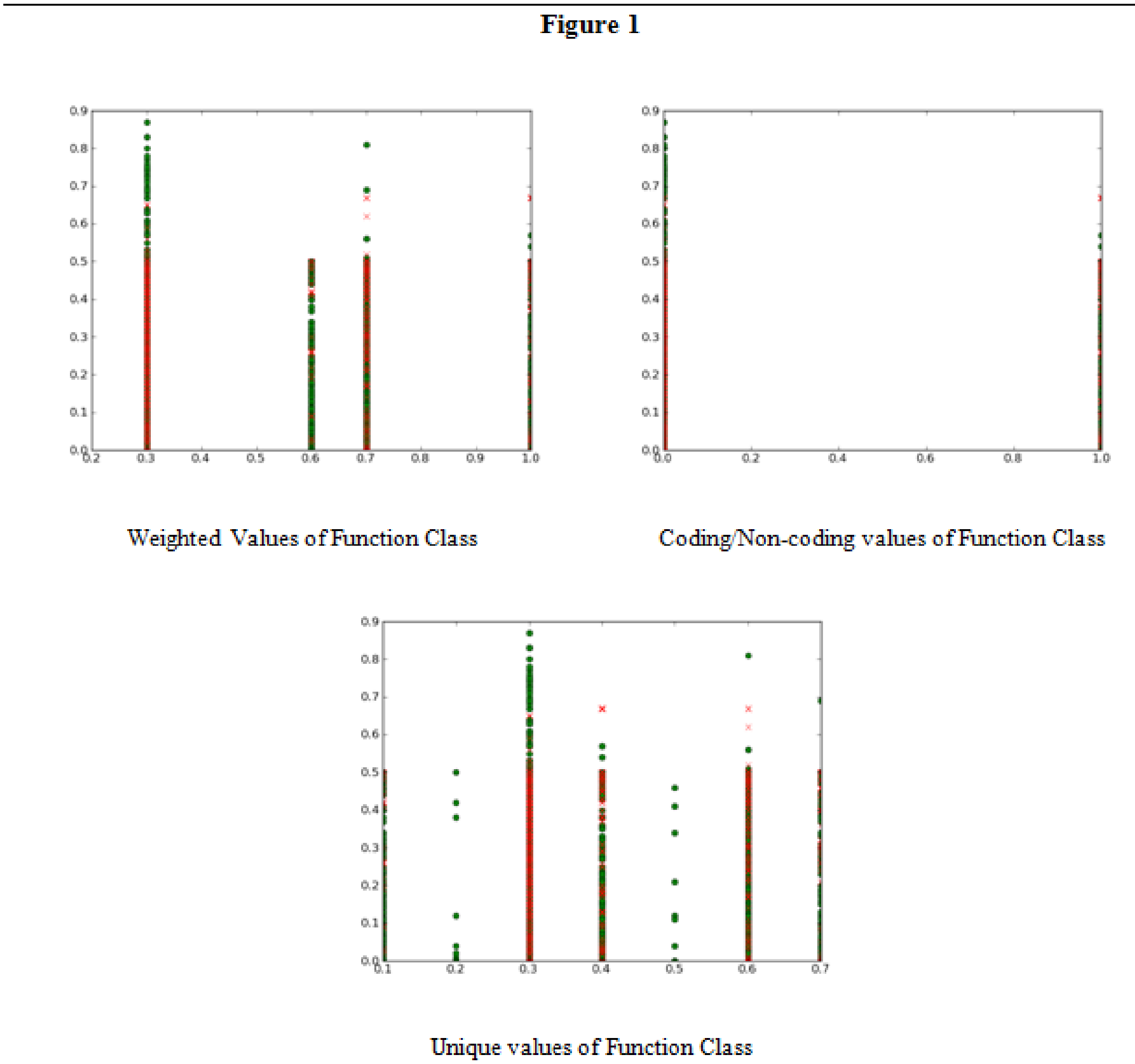


Figure 1

Weighted Values of Function Class

Coding/Non-coding values of Function Class

Unique values of Function Class

Fig. 10, Data normalization result, Result displaying the nonlinear result obtained after normalization of disease and non-disease class against Function class attribute of SNP.

CHAPTER FIVE

DISCUSSION

5.1 Significance of cytokines and their role in diseases

The role of cytokine in diseases can be understood by the realization of their important role they play in immune response of body. They regulate the immune response of the body and control the innate and adaptive response of the body. This is very significant in the sense that significant numbers of pathways that are involved in B cell and T cell regulation are mediated by the mechanism of cytokine action. This is underscored by the several literature studies which have implicated the association of cytokines to diseases. The knowledge of cytokine genes association to diseases as the database suggest can be an important repertoire to understand the immunology behind genetic causes of diseases. This can benefit the immunogical and biology community to understand and illuminate the pathways involved in the immunology of diseases. The knowledge of biological pathways associated with diseases can provide clues on causes of diseases and better targets for drug action and for further concentrating on the genetic element which can also be a cause of diseases or potential cause of diseases. The potential mechanism of drug action is mostly on the immunological pathways and sometimes genetic elements; improving on this knowledge can significantly improve on understanding the diseases and body defense mechanism against them and can help in improving the defense mechanism of the body.

5.1.1 Significance of the findings of the cytokine SNPs association to diseases

Several genome wide association studies have implicated the association of SNPs to diseases. As the database disease data suggest there are 1737 distinct SNPs associated to diseases. These SNPs can act as the biomarker for diseases or can be potential cause of diseases. As, a biomarker they can help target the genetic element more correctly. With a linkage disequilibrium study, they can also implicate the association of other genetic elements to diseases. This can further aid in understanding and building the pathways of the genes associated to diseases and hence providing broad and probably better understanding of the genetics behind immunological action against diseases. SNPs can also causally be associated with diseases if they are mutations which effect the expression of the amino acids and hence disordered proteins or proteins with different biological functions which result in causing diseases. These mutations can be potential and probably better targets of drug action as they are specific, vary from population to population, localized and indicate distinctively known causes of diseases. This knowledge can help in designing and defining better drugs for diseases; it can also help in providing solution for tailor made medicine and personalized medicine. SNPs associated to diseases, particularly non-Synonymous SNPs can hence help in improving the drug design, defining drug target, target validation, improving drug action with increasing specificity and uniqueness of drug action.

5.1.2 Characteristic of SNP presence in the Cytokine genetic region associated with disease

SNPs present in the genetic element of diseases are specific marker for the haplotype which can implicate the genetic elements to be associated with diseases or specifically a particular gene to be associated with diseases. This can provide information on the association of genetic elements to be associated with disease or the evolutionary mechanism behind the causal association of haplotype to diseases. The SNP alleles can characteristically be associated with diseases of specific population groups if found to be associated with diseases and can even be more specific biomarker for diseases and hence the suitable and better site for drug action. This characterizes how the specificity increased by the knowledge of SNP can increase the drug selection and drug action; potentially also this suggest a mechanism for the knowledge of new drug regimens and with specific disease target known and its causality or association biologically validated.

5.1.3 Purpose of the findings and application of the database

DACS-DB stores cytokines associated with diseases and specifically Cytokine SNPs associated with diseases. This includes the vital information of cytokine SNPs associated to more lethal and prevalent diseases such as cancers, heart diseases, auto immune diseases, brain related diseases(Parkinson disease, Alzheimer's disease, ADHD, Bipolar diseases), skin diseases, eye diseases etc. This can provide the potential drug targets for diseases. This information can help in understanding the cytokines associated with disease pathways and can help in constructing the disease pathways. They can also help in building the immune pathways and the cytokines associated can provide vital information on understanding and deciphering the innate and adaptive immune response pathways and the mechanism of action of B and T cell and the immune regulatory mechanism of disease associated pathways.

The database can be applicable in providing information about cytokines associated to diseases, SNPs associated to diseases, diseases associated to cytokines and diseases associated to cytokine SNPs. This can be vital in understanding the biomarker association of SNPs to genetic elements and also providing causal association of SNPs to disease. This can be very important information with regards to drug design, drug discovery, target discovery, target validation and individualized drug design and personalized medicine discovery. In the future, there is a plan to build a disease association potential prediction tool which can provide information on disease association potential for newly discovered SNPs and hence a new avenue for disease target association and prediction.

5.2 Importance of DACS-DB database

The importance of DACS-DB database can be underscored by the role of SNP information that can be helpful in determining the role of causality of diseases and the role of SNP as biomarker to find the genetic association of SNP to diseases. This can be helpful in determining whether the diseases can be associated to cytokine SNP. This information can be crucial in understanding the role of genetic causes of diseases and whether this knowledge can be helpful in understanding the genetic causes of diseases and this can be crucial in understanding the role B-cells and T-cells play in determining the immune system of the body. Cytokine knowledge can also be helpful in understanding the role of growth hormones in regulating the growth, maintenance and development of the body cells. This play important roles in regulating the body homeostasis and proper functioning and regulation of different biological processes of the body. There are important cytokines which regulate the innate and adaptive immune response of the body, and this knowledge can be crucial in determining the role of Immune system and disease interaction and the role it can play role in drug target determination and ascertaining the role of the immune system and its associated pathways in drug mediated disease regulation. The DACS-DB database can be crucial in assigning the importance of disease causal association to genetic causes and hence can be a valuable information source for biomedical researchers, immunologists and biologists.

5.3 Role of DACS-DB in suggesting solution for genetic treatment

Since the DACS-DB database is a collection of Cytokine SNPs with important diseases associated attribute information and disease information along with SNP attributes, and it can prove to be a guide for important information on cytokine SNP biomarkers associated with diseases. With chromosomal location of a SNP known it can provide information on linkage disequilibrium associated SNP associated to a particular SNP which can give information on haplotype associated to a disease which also be a suitable biomarker and provide information on association of SNP to a genetic element or genetic elements and their association to diseases. The database has information on genetic function class which can provide information on whether a SNP can causally be associated to a disease. If a SNP is causally associated to disease it can be a suitable drug target as changing it or affecting it can have a positive effect on curing disease. SNP as biomarker can provide information on genetic elements associated to diseases as these genetic elements can be suitable drug targets and can assist in drug design and checking existing drugs whether they can affect the existing disease condition or not. Most of the diseases have genetic elements and SNPs associated to diseases and they can be a suitable drug target for disease treatment and drug development and play important roles in pharmacogenetics analysis of disease. SNPs as biomarker can play important role in SNP profiling information and SNP composition information of a population group or individual and can provide biomarker for drug effect on the particular population group and whether the same drug can be suitable for another population group if they have similar genetic profiling. The disease information associated with genetic elements can provide information on the genetic cause of diseases and how effecting these genetic elements can have effects on treatment of diseases.   Disease associated to Synonymous SNPs (Missense, frame shift, non-sense function class) can have information on genetic causality of disease or diseases while Synonymous SNPs can be very suitable biomarker for genetic assay and can provide genetic association to diseases which can be important in genetic treatment of diseases. Most of the diseases particularly immunological compatible one for which cytokines and their SNPs are known , can genetically be treated and also, they can be monitored beforehand and prevented before they occur, by modifying the disease causing mutation or administering drug to correct the genetic causes which can cause it to happen in later stage of life particularly for the individuals and also for the population groups.

CHAPTER SIX

CONCLUSION

The DACS-DB has been designed, developed and published on the web as a user-friendly collection of disease associated cytokine gene SNPs. Thus, we present the first catalog of highly annotated cytokine SNP data that is searchable for researchers studying genomic variations in diseases. Various types of queries can be performed to gain knowledge on cytokines implicated in diseases. Such knowledge pertaining to the insinuation of cytokine gene variations in diseases can be useful to immunologists, biomedical researchers, pharmaceutical companies etc. and may help understand the genetics behind the diseases thus aiding in next generation drug development.

REFERENCES:

1.) Bidwell J., 1999, Cytokine gene polymorphism in human disease: on-line databases, Genes and Immunity (1999) 1, 3–19

2.) Wicher J. T., 1990, Cytokines in diseases, CLIN. CHEM., 36/7, 1269-1281

3.) P. K. Gupta, J. K. Roy and M. Prasad, 2001, Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants, CURRENT SCIENCE, VOL. 80, NO. 4, 25 FEBRUARY 2001

4.) Anthony J. Brookes, 1999, The essence of SNPs, Gene Volume 234, Issue 2, 8 July 1999, Pages 177-186

5.) Francis S. Collins,1 Lisa D. Brooks and Aravinda Chakravarti ,1998, A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation, Genome Res. 1998. 8: 1229-1231

6.) Kwok, P.-Y., Q. Deng, H. Zakeri, S.L. Taylor, and D.A. Nickerson. 1996. *Genomics* **31:** 123–126.

7.) Bell, D.A., Taylor, J.A., Butler, M.A., Stephens, E.A., Wiest, J., Brubaker, L.H., Kadlubar, F.F., Lucier, G.W., 1993, Genotype/phenotype discordance for human arylamine N-acetyltransferase (NAT2) reveals a new slow-acetylator allele common in African-Americans. Carcinogenesis 14, 1689– 1692.

8.) Charles A. Dinarello et. al., Proinflammatory Cytokines 2000)*,Chest* 2000;118;503-508.

9.) Steven M. Opal et. al., 2000, Anti-Inflammatory Cytokines, *Chest* 2000;117;1162-1172

10.) Stavros C. Manolagas et. al., 1995, Bone Marrow, Cytokines, and Bone Remodeling — Emerging Insights into the Pathophysiology of Osteoporosis, Volume 332:305-311

11.) Jan Vil č ek et. al., 2004, Historical review: Cytokines as therapeutics and targets of therapeutics

12.) S. SKURKOVICH et. al.,1987, A Unifying Model of the Immunoregulatory Role of the Interferon System: Can Interferon Produce Disease in Humans, CLINICAL IMMUNOLOGY AND IMMUNOPATHOLOGY 43, 362-373

13.)CJ Sanderson et. al., 1992, Interleukin-5, eosinophils, and disease, Blood**,** 1992 79: 3101-3109

14.)Charles A. Dinarello et. al., 1993*,* The Role of Interleukin-1 in Disease, The new England journal of medicine, Volume 328:106-113

15.)Dimitrias A. Papanicolaou et. al. , 1998, The Pathophysiologic Roles of Interleukin-6 in Human Disease, Ann Intern Med*.* 1998; 128:127-137.

16.)Peter J. Mannon et. al. , 2004, Anti–Interleukin-12 Antibody for Active Crohn's Disease, N Engl J Med 2004;351:2069-79.

17.) Kevin, J Tracey et. al. , 1993, Tumor necrosis factors other cytokines and diseases, Annu. Rev. Cell Bio., 1993, 317-343

18.) Warren J. Leonard et. al. , 2001, CYTOKINES AND IMMUNODEFICIENCY DISEASES, Nature Immunology*,* 200 | DECEMBER 2001 | VOLUME 1

19.) Konstantinos A. Papadakis et. al. , 2000, ROLE OF CYTOKINES IN THE PATHOGENESIS OF INFLAMMATORY BOWEL DISEASE, Annu. Rev. Med. 2000. 51:289–298

20.) Sergio Romagnani et. al. , 1994, L YMPHOKINE PRODUCTION BY HUMAN T CELLS IN DISEASE STATES, Annu. Rev. Immunol. 1994. 12:227-57

21.) Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JAE *et al, 2009,* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet* 2009, 41(9):1006-1010.

22.) PR Burton, DG Clayton, LR Cardon, N Craddock,2007, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007, 447(7145):661-678.

23.) Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, Bailey R, Bourget K, Plagnol V, Field S, Atkinson M *et al, 2007,* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat Genet* 2007, 39(9):1074-1082.

24.) Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, Yang JH, Howson JM, Stevens H, McManus R, Wijmenga C *et al,2008,* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N Engl J Med* 2008, 359(26):2767-2777.

25.) Garcia VE, Chang M, Brandon R, Li Y, Matsunami N, Callis-Duffin KP, Civello D, Rowland CM, Bui N, Catanese JJ *et al, 2008* Detailed genetic characterization of the interleukin-23 receptor in psoriasis. *Genes Immun* 2008, 9(6):546-555.

26.) Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST *et al,2005,* Complement factor H polymorphism in age-related macular degeneration. *Science* 2005, 308(5720):385-389.

27.) Dendrou CA, Plagnol V, Fung E, Yang JHM, Downes K, Cooper JD, Nutland S, Coleman G, Himsworth M, Hardy M *et al,2005,* Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat Genet* 2009, 41(9):1011-1015.

28.) Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K,2001, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001, 29(1):308-311.

29.) Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L *et al,2001,* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007, 39(10):1181-1186.

30.) Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE,2002, PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002, 30(1):163-165.

31.)Ju-Youn Lee, InSong Koh,2001, Drug to SNP: A Pharmacogenomics Database for Linking Drug Response to SNPs, Genome Informatics 12: 482–483 (2001)

32.)P.H.Lee et. al., 2008, F-SNP: computationally predicted functional SNPs for disease association studies, Nucl. Acids Res. (2008) 36 (suppl 1): D820-D824.

33.)T. A. Manolio et. al., 2008, A HapMap harvest of insights into the genetics of common disease, J Clin Invest. 2008 May 1; 118(5): 1590–1605.

34.)B.R. Packer et. al., 2004, SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. Nucl. Acids Res. (2004) 32 (suppl 1): D528-D532.

35.)D. Fredman et.al., 2002, HGVBASE: a human sequence variation database emphasizing data quality and broad spectrum, Nucl. Acids Res. (2002) 30 (1): 387-391.

36.)J. Bidwell et.al., 2001, Cytokine gene polymorphism in human disease: on-line databases, Genes and Immunity (2001), Supplement 1, 2, 61-70

37.)P. Yue et.al., 2006, SNPs3D: Candidate gene and SNP selection for association studies, *BMC Bioinformatics* 2006, **7:**166

38.)Y. Zhang et.al., 2010, Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information**,** BMC Med Genomics. 2010; 3: 1

APPENDICES:

Supplement 1: PHP codes of SNP search page***.***

```
<HTML>

<head><title>SNP Search </title></head>

<BODY>

<h1><a href=index.php>Home</a><h1>

<?php

echo("<h2><center>SNP search Information</center></h2>");

echo("<hr>");

/*echo $_POST['searchname'];*/

/*echo $_REQUEST['searchname'];*/

    import_request_variables('p','p_');

    $db= mysql_connect("libra45.uits.iu.edu:3234","root","dis_snp10") or die(mysql_error());

    mysql_select_db('dacsdb',$db)or die("Cannot select database.");

    $result=mysql_query("SELECT * FROM SNP WHERE $p_searchtype like '$p_searchname'");

    $result1=mysql_query("SELECT GeneId FROM SNP WHERE $p_searchtype like '%$p_searchname%'");

    $result2=mysql_query("SELECT GeneSymbol FROM SNP WHERE $p_searchtype like '%$p_searchname%'");

    $result3=mysql_query("SELECT Chromosome FROM SNP WHERE $p_searchtype like '%$p_searchname%'");

    $result4=mysql_query("SELECT GeneSymbol,GeneId,SNPId,ChromosomalLocation,FunctionClass,Allele,HeterozygosityValue,HetStdErr,5primeSequence,3primeSequence FROM SNP WHERE $p_searchtype='$p_searchname'");

    $columnnos=mysql_num_fields($result4);

    print "<table cellpadding=0 cellspacing=0 width=100% border=1 >";

    for($i=0;$i<$columnnos;$i++)

        {

            $attribute_name=mysql_field_name($result4,$i);

            if($p_searchtype=='Chromosome')
            {

    //if($attribute_name=='Chromosome'||$attribute_name=='GeneSymbol'||$attribute_name=='GeneId'||$attribute_name=='SNPId'||$attribute_name=='ChromosomalLocation'||$attribute_name=='FunctionClass'||$attribute_name=='Allele'||$attribute_name=='HeterozygosityValue'||$attribute_name=='HetStdErr'){

            print "<td align=left><b>".$attribute_name."</b></td>";

            }

            elseif($attribute_name=='Sequence5'||$attribute_name=='Sequence3')

            {

            print "<td align=left><b>".$attribute_name.'\''."</b></td>";

            }

        }

    //print "$p_searchtype";

    //print "'$p_searchname'";

            if($p_searchtype=='Chromosome')

            {//print "'$p_searchtype'";

            $total=0;

    while ( $tuple=mysql_fetch_array($result4) ) /* print data of each tuple that meets the search query criteria */

        {

        print "<tr>";

        $total++;

            for ( $i=0;$i<$columnnos;$i++ )

            {

            $attribute_value=$tuple[$i];

            print "<td align=left>".$attribute_value . "</td>";


            }

        }
    }
    mysql_free_result($result4);

    if($p_searchtype=='GeneSymbol'||$p_searchtype=='GeneId'||$p_searchtype=='SNPId'||$p_se
archtype=='FunctionClass')
    {
     /*echo $p_searchname;*/
     /*echo $p_searchtype;*/
        $ncols=mysql_num_fields($result);/* How many columns (fields) */
        print "<table cellpadding=0 cellspacing=0 width=100% border=1 >";
        print "<tr>";
        for ( $i=0; $i <$ncols; $i++ )
        {
        if($i==0||$i==1||$i==2){next;}
        else{

        $column_name=mysql_field_name($result,$i);   /* field name */
            print "<td align=left><b>".$column_name."</b></td>";
        }}
        print "</tr>";
    $total=0;
    while ( $row = mysql_fetch_array($result) ) /* process result row-by-row */
    {
        print "<tr>";
        $total++;
        for ( $i=0; $i <$ncols; $i++ )
        {
        if($i==0||$i==1||$i==2){next;}
        else{
            $column_value=$row[$i];          /* field value */
            print "<td align=left>".$column_value . "</td>";


        }}
    }
    }
    mysql_free_result($result);
    echo ("<form method=get action=cleanData1.php>");
    echo ("<form method=get action=cleanDataCs.php>");
    if($p_searchtype==GeneId){
    $row1 = mysql_fetch_array($result1);$row3 = mysql_fetch_array($result3);$row2 = 
mysql_fetch_array($result2);
    echo ("<h3>Search for ".$p_searchtype ." ".$p_searchname." "."GeneSymbol"." ".$row2[0]." 
"."located on chromosome"." ".$row3[0].  ": ". $total." records found</h3>");}
    elseif($p_searchtype==GeneSymbol){
    $row1 = mysql_fetch_array($result1);$row3 = mysql_fetch_array($result3);$row2 = 
mysql_fetch_array($result2);
    echo ("<h3>Search for ".$p_searchtype ." ".$p_searchname." "."GeneId"." ".$row1[0]." 
"."located on chromosome"." ".$row3[0].  ": ". $total." records found</h3>");
    echo("<p align=\"left\"><a 
href=\"cleanData1.php?searchtype=$p_searchtype&searchval=$p_searchname\">Click here to 
download</a></p>");
    }
    elseif($p_searchtype==Chromosome){
    //$row1 = mysql_fetch_array($result1);$row3 = mysql_fetch_array($result3);$row2 = 
mysql_fetch_array($result2);
    echo ("<h3>Search for ".$p_searchtype ." ".$p_searchname.  ": ". $total." records 
found</h3>");echo("<p align=\"left\"><a 
href=\"cleanDataCs.php?searchtype=$p_searchtype&searchval=$p_searchname\">Click here to 
download</a></p>");}//" "."GeneId"." ".$row1[0]." "."located on chromosome"." ".$row3[0].  ": ". 
$total." records found</h3>");}
```

```
        elseif($p_searchtype==FunctionClass){

        echo ("<h3>Search for ".$p_searchtype ." ".$p_searchname.  ": ". $total." records
found</h3>");}//" "."GeneId"." ".$row1[0]." "."located on chromosome"." ".$row3[0].  ": ". $total."
records found</h3>");}

        elseif($p_searchtype==SNPId){

        echo ("<h3>Search for ".$p_searchtype ." ".$p_searchname.  ": ". $total." records
found</h3>");}//" "."GeneId"." ".$row1[0]." "."located on chromosome"." ".$row3[0].  ": ". $total."
records found</h3>");}

        //echo ("<br><left><h1><a href=www.google.com> Google </a></h1></center>");

        //echo ("<br><left><h1><a href=Cytokine_SNP.php> Back</a></h1></center>");

        echo ("<hr>");

?>

</BODY>

</HTML>
```

Supplement 2 Disease search PHP code

```
<?php
echo("<HTML><head><head><title>Disease Information </title></head>");
echo("<BODY>");
echo("<h1><a href=index.php>Home</a></h1>");//echo ("<br><left><h1><p align=\"right\"><a
href=Cytokine_associated_diseases.php> Back</a></h1></center>");
echo("<h2><left>Disease Information</left></h2>");
echo("<hr>");
/*echo $_POST['searchname'];*/
/*echo $_REQUEST['searchname'];*/
        import_request_variables('p','p_');
        $db= mysql_connect("libra45.uits.iu.edu:3234","root","dis_snp10") or die(mysql_error());
        mysql_select_db('dacsdb',$db)or die("Cannot select database.");
        $result=mysql_query("SELECT *  FROM Disease WHERE $p_searchtype like
'%$p_searchname%'");
        $result1=mysql_query("SELECT Disease,SNPId,link,Year,PMID  FROM Disease WHERE
$p_searchtype like '%$p_searchname%'");//='$p_searchname'");//
        $result2=mysql_query("SELECT GeneId,GeneSymbol,Disease,SNPId,link,Year,PMID  FROM
Disease WHERE $p_searchtype like '%$p_searchname%'");
        $result3=mysql_query("SELECT GeneId,GeneSymbol,Disease,SNPId,link,Year,PMID  FROM
Disease WHERE $p_searchtype like '%$p_searchname%'");
        $result4=mysql_query("SELECT GeneId,GeneSymbol,Disease,SNPId,link,Year,PMID  FROM
Disease WHERE $p_searchtype like '%$p_searchname%'");
        //mysql_free_result($SNP);
                /*echo $p_searchname;*/
         /*echo $p_searchtype;*/
                if($p_searchtype=='GeneId'||$p_searchtype=='GeneSymbol')
                {
                $columnnos=mysql_num_fields($result1);
                print "<table cellpadding=0 cellspacing=0 width=40% border=1 >";
                print "<tr>";


                for ( $i=0; $i <$columnnos; $i++ )
                {
                if($i==2){next;}else{
                $column_name=mysql_field_name($result1,$i); /* field name */
                        print "<td align=left><b>".$column_name."</b></td>";
            }}}
                print "</tr>";
                $total=0;
        while ( $tuple = mysql_fetch_array($result1) ) /* process result row-by-row */
        {
                print "<tr>";
                $total++;
                for ( $i=0; $i <($columnnos); $i++ )
                {
                        if($i<($columnnos-1))
                        {if($i==2){next;}else{
                        $column_value=$tuple[$i];
                                print "<td align=left>".$column_value."</td>";}
                        }
                        elseif($i==3)
                        {
                        //while ( $tuple1 = mysql_fetch_array($SNP) )
                        echo"$tuple1";
                                //{$column_value1=$tuple1;}
                        $column_value=$tuple[3];          /* field value */
                        $column_value1=$tuple1[3];
```

```
            if($column_value=$column_value1)

            {print "<td align=left><a target=\"_blank\"
href='$column_value'>".$tuple1."</a></td>";}

            }



            elseif($i==4)

            {

                $column_value1=$tuple[4];

            $column_value=$tuple[2];        /* field value */

            print "<td align=left><a target=\"_blank\"
href='$column_value'>".$column_value1."</a></td>";

            }

        }

    }

    mysql_free_result($result1);

    echo ("<form method=get action=cleanDataGs.php>");

    echo ("<form method=get action=cleanDataGi.php>");

    if($p_searchtype==GeneId){

    echo ("<h3>Search for " .$p_searchtype ." ".$p_searchname.": ". $total." records
found</h3><br>");

    echo("<p align=\"left\"><a
href=\"cleanDataGi.php?searchtype=$p_searchtype&searchval=$p_searchname\">Click here to
download</a></p>");}

    //echo("<p align=\"right\"><a href=\"cleanData.php?searchtype=GeneId\">Click here to
download</a></p>");}



    if ($p_searchtype==GeneSymbol){

        echo ("<h3>Search for " .$p_searchtype ." ".$p_searchname.": ". $total." records
found</h3>");
```

```
                    echo("<p align=\"left\"><a
href=\"cleanDataGs.php?searchtype=$p_searchtype&searchval=$p_searchname\">Click here to
download</a></p>");

}


          if($p_searchtype==Disease)

          {

          $ncols=mysql_num_fields($result);/* How many columns (fields) */

          print "<table cellpadding=0 cellspacing=0 width=70% border=1 >";

          print "<tr>";

          for ( $i=0; $i <$ncols; $i++ )

                    {

                    if($i==4)

                              {next;}

                    else

                              {

                              $column_name=mysql_field_name($result,$i);   /* field name */

                              print "<td align=left><b>".$column_name."</b></td>";

                    }

                    }

           }

                    print "</tr>";

          $total=0;

          //$ncols2=mysql_num_fields($SNPunique)

 //while ( $row2 = mysql_fetch_array($SNPunique) )

          //{ echo "$row2"; }


 while ( $row = mysql_fetch_array($result) ) /* process result row-by-row */
```

```
          {
                    print "<tr>";
                    $total++;
                    for ( $i=0; $i <=($ncols); $i++ )
                              {
                                        if($i<($ncols-1))
                                        {
                                        if($i==4)
                                                  {next;}
                                        else
                                                  {
                                                  $column_value=$row[$i];
                                                  print "<td align=left>".$column_value."</td>";
                                                  }
                                        }
                                        elseif($i==6)
                                        {
                                        $column_value1=$row[6];
                                        $column_value=$row[4];        /* field value */
                                        print "<td align=left><a target=\"_blank\"
href='$column_value'>".$column_value1."</a></td>";
                                        }
                              }
           }
          mysql_free_result($result);
echo ("<form method=get action=cleanData.php>");
if($p_searchtype==Disease)
          {
```

```
        echo ("<h3>Search for " .$p_searchtype ." ".$p_searchname.": ". $total." records found</h3>");

    echo("<p align=\"left\"><a
href=\"cleanData.php?searchtype=$p_searchtype&searchval=$p_searchname\">Click here to
download</a></p>");

        }

if($p_searchtype==SNPId)

                {

                $ncols=mysql_num_fields($result2);/* How many columns (fields) */

                print "<table cellpadding=0 cellspacing=0 width=50% border=1 >";

                print "<tr>";

                for ( $i=0; $i <$ncols; $i++ )

                {

                if($i==3||$i==4){next;}else{

                $column_name=mysql_field_name($result2,$i); /* field name */

                        print "<td align=left><b>".$column_name."</b></td>";

            }}}

                print "</tr>";

        $total=0;

if($p_searchtype==SNPId)

{

        while ( $row = mysql_fetch_array($result2) ) /* process result row-by-row */

        {

                print "<tr>";

                $total++;

                for ( $i=0; $i <($ncols); $i++ )

                {

                        if($i<($ncols-1))

                        {if($i==3||$i==4){next;}else{
                        $column_value=$row[$i];

                                  print "<td align=left>".$column_value."</td>";}

                        }

                        elseif($i==6)

                        {

                                  $column_value1=$row[6];

                        $column_value=$row[4];           /* field value */

                        print "<td align=left><a target=\"_blank\"
href='$column_value'>".$column_value1."</a></td>";

                        }

                        }

        }

}

        mysql_free_result($result2);

if($p_searchtype==SNPId)

{

        echo ("<h3>Search for " .$p_searchtype ." ".$p_searchname.": ". $total." records found</h3>");

}

if($p_searchtype==Year)

                {

                $ncols=mysql_num_fields($result3);/* How many columns (fields) */

                print "<table cellpadding=0 cellspacing=0 width=50% border=1 >";

                print "<tr>";

                for ( $i=0; $i <$ncols; $i++ )

                {

                if($i==4||$i==5){next;}else{

                $column_name=mysql_field_name($result3,$i); /* field name */

                        print "<td align=left><b>".$column_name."</b></td>";
```

```
      }}}
        print "</tr>";
    $total=0;
if($p_searchtype==Year)
{
    while ( $row = mysql_fetch_array($result3) ) /* process result row-by-row */
    {
        print "<tr>";
        $total++;
        for ( $i=0; $i <($ncols); $i++ )
        {
            if($i<($ncols-1))
            {if($i==4||$i==5){next;}else{
            $column_value=$row[$i];
                print "<td align=left>".$column_value."</td>";}
            }
            elseif($i==6)
            {
                $column_value1=$row[6];
            $column_value=$row[4];      /* field value */
            print "<td align=left><a target=\"_blank\"
href='$column_value'>".$column_value1."</a></td>";
            }
            }
    }
}
    mysql_free_result($result3);
if($p_searchtype==Year){
    echo ("<h3>Search for " .$p_searchtype ." ".$p_searchname.": ". $total." records found</h3>");}
if($p_searchtype==PMID)
        {
            $ncols=mysql_num_fields($result4);/* How many columns (fields) */
            print "<table cellpadding=0 cellspacing=0 width=50% border=1 >";
            print "<tr>";
            for ( $i=0; $i <$ncols; $i++ )
                {
                    if($i==4||$i==5)
                        {next;}
                    else
                        {
                        $column_name=mysql_field_name($result4,$i); /* field
name */
                        print "<td align=left><b>".$column_name."</b></td>";
                    }
                }
        }
        print "</tr>";
        $total=0;
if($p_searchtype==PMID)
{
        while ( $row = mysql_fetch_array($result4) ) /* process result row-by-row */
            {
            print "<tr>";
            $total++;
            for ( $i=0; $i <($ncols); $i++ )
                {
```

```
                if($i<($ncols-1))
                    {
                    if($i==4||$i==5)
                        {next;}
                    else
                    {
                    $column_value=$row[$i];
                    print "<td align=left>".$column_value."</td>";
                    }
                    }
                elseif($i==6)
                    {
                    $column_value1=$row[6];
                    $column_value=$row[4];        /* field value */
                    print "<td align=left><a target=\"_blank\"
href='$column_value'>".$column_value1."</a></td>";
                    }
                }
            }
}
        mysql_free_result($result4);
            if($p_searchtype==PMID)
                {
                    echo ("<h3>Search for " .$p_searchtype ." ".$p_searchname.":
". $total." records found</h3>");
                }
//echo ("<br><left><h1><a href=Cytokine_associated_diseases.php> Back</a></h1></center>");
```

```
//	# Original PHP code by Chirp Internet: www.chirp.com.au
//  # Please acknowledge use of this code by including this header.
?>
</BODY>
</HTML>
```

Supplement 3 PHP code for downloading SNP data

```
<?php
function cleanData(&$str)
  {
    $str = preg_replace("/\t/", "\\t", $str);
    $str = preg_replace("/\r?\n/", "\\n", $str);
    if(strstr($str, '"')) $str = '"' . str_replace('"', '""', $str) . '"';
  }
//import_request_variables('p','p_');
//  # file name for download
  $filename = "website_data_" . date('Ymd') . ".xls";

//header('Content-Disposition: attachment; filename="downloaded.pdf"');
$searchname=$_GET["searchval"];
$searchtype=$_GET["searchtype"];
header("Content-Disposition:attachment;filename=\"$filename\"");
header("Content-Type:application/vnd.ms-excel");

// Displaying Information.
  $flag = false;
  import_request_variables('p','p_');
  $db= mysql_connect("libra45.uits.iu.edu:3234","root","dis_snp10") or die(mysql_error());
  mysql_select_db('dacsdb',$db)or die("Cannot select database.");

  //$result = mysql_query("SELECT * FROM Disease ORDER BY GeneSymbol") or die('Query failed!');
    $result = mysql_query("SELECT GeneId,GeneSymbol,Disease,SNPId,Year,PMID  FROM Disease WHERE $searchtype like '%$searchname%'") or die('Query failed!');
```

```
 while(false !== ($row = mysql_fetch_assoc($result))) {
  if(!$flag) {
   # display field/column names as first row
   echo implode("\t", array_keys($row)) . "\n";


        //echo("something unimportant")."\n";
   $flag = true;
  }
  array_walk($row, 'cleanData');
  echo implode("\t", array_values($row)) . "\n";
 }
?>
```

}

Supplement 4: Perl code to retrieve SNP data.

Perl code for parsing data from XML file:

```
#!/usr/bin/perl

$output="SNP.txt";

open(FILE, "412SNPfile.txt") or die("Unable to open file");

open(OUTPUT, ">$output") or die("Unable to open file");

@array=<FILE>;

$ref=0;

foreach $array (@array)

{

if($array=~/\<Rs\srsId\=\"(.*)\"\ssnpClass/)

{

$i=0;

$j=0;

print OUTPUT "\n\n$1\t";

}

if($array=~/\<Het\stype\=.*value\=\"(.*)\"\sstdError\=\"(.*)\>/)

{

if($1ne''&$2ne'')

{

print OUTPUT "$1\t$2\t";

}

else

{

print "null \t null";

}

}
```

```
if($array=~/\<Validation\s(.*)\>\<\/Validation\>/ || $array=~/\<Validation\s(.*)\>/ )
{
if($1ne'')
{
print OUTPUT "$1\t";
}
else
{
print "null";
}
}
if($array =~/\<Ss\sssId.*/)
       {
        $ref_flag=0;
        next;
       }
     if ($array=~/\<Sequence\sexemplar.*/)
       {
        $ref_flag=1;
       }
      if($ref_flag ==1)
       {
          if($array=~/\<Sequence\sexemplar.*/)
          {


          }
          else
          {
```

```
        $array=~s/\s+//g;
        $array=~s/\n//g;
        $array=~s/\<Seq5\>(.*)\<\/Seq5\>/$1/g;
        $array=~s/\<Seq3\>(.*)\<\/Seq3\>/$1/g;
        $array=~s/\<Observed\>(.*)\<\/Observed\>/$1/g;
        $array=~s/\<\/Sequence\>//g;
        print OUTPUT "$array\t";


    }
        }
if($array=~/\<Assembly\sdbSnpBuild\=\"(.*)\"\sgenomeBuild.*Label\=\"Celera\".*/)
{
print OUTPUT "$1\t";
}
if($array=~/\<Component\scomponentType.*chromosome\=\"(.*)\"\sstart.*Celera.*/)
{
print OUTPUT "$1\t";
}
if($array=~/\<MapLoc.*orient\=\"(.*)\"\sphysMapInt\=\"(.*)\"\sleftFlank.*/g)
{
$j++;
if($j==1)
{
print OUTPUT "$2\t";
}
}
if($array=~/\<FxnSet\sgeneId\=\"(.*)\"\ssymbol\=\"(.*)\".*\sfxnClass\=\"(.*)\>/)
{
```

```
$i++;

if($i==1)

{

print OUTPUT "$1\t$2\t$3";

}

}

}
```

Sushant Bhushan

719, Indiana Avenue, Ste 319, Indianapolis, IN 46202
Phone (317) 522 – 6531, **Email: sbhushan@iupui.edu**

- Master of Science in Bioinformatics (August, 2011)
- 5 year experience in Perl programming and 3 year experience in MySQL and PHP programming
- Expertise in database design, development and management
- Experience in Web design and development
- Experience in software installation such as Partek, MEME and working in UNIX environment
- 6 months Hands on experiences with high throughput genomic data

OBJECTIVE: Seeking a full–time position in Bioinformatics which involves software development, database development or data analysis.

EDUCATION:

- **M.S, Bioinformatics**, Indiana University, School of Informatics, Indianapolis
    - GPA – **(3.4/4.0),** June 2011
- **B-Tech, Biotechnology**, Visvesvaraya Technological University, Belgaum, India
    - GPA – **(3.4/4.0),** May 2006.

SKILL SET:

Information Technology:

- Languages & Scripts: JAVA, Perl, Python, Ruby on Rails, R, Bioconductor, MySQL, HTML, XML, CGI, C.
- Platforms: UNIX, LINUX, Macintosh, Windows XP/2000/Vista/7.
- Statistical tools: SPSS, SAS
- Tools: dbSNP, RefSeq, Hap Map, UCSC Genome Browser, Ensemble, Blast, MEME, TRANSFAC, Gene GO, Clustal W, Modeller9.1, Spdbv, PHYLIP, PAUP
- Pathway building tool : MetaCore
- Database: MySQL, SQL, Oracle

RESEARCH WORK EXPERIENCE:

INDIANA UNIVERSITY:

- **Graduate Research Assistant** – Indiana University School of Informatics, Indianapolis (May 09 – Present)
    - As part of thesis work developed the DACS-DB database available online at www.iupui.edu/~cytosnp and worked on to classify SNP data into disease class and non-disease class.

- **Graduate Teaching Assistant**– Dr. Narayanan B. Perumal, Introduction to Informatics (I-501 course), Indiana University School of Informatics, Indianapolis (January 2010 – May 2010)
- **Scientific Programmer** – CCBB, IU School of medicine, Indianapolis (Sept 2010 – Oct 2011)
    - High through put data analysis. Worked on ChIP-Seq data analysis for determining association between promoter regions and H3K4Me3 peaks
- **Lecturer** – KLE College of engineering and technology, Belgaum, India, (Aug 2007 – May 2008)

ACADEMIC & COURSE PROJECTS:

- **Machine Learning and Pattern Recognition:**
    - "*Classification of Cytokine SNP data with SVM*" – Application of SVM method for classification of disease and non-disease data of Cytokine SNPs. Application of RBF, MLP, and Perceptron was previously applied to test and train the SNP data into disease and non-disease class.
- **Translational Bioinformatics**:
    - "Performed the class Analysis of protein structure data to determine the docking potential of proteins and drugs with docking and comparing it to machine learning approach to determine which approach is better for screening better drug to interact to protein.
- **Introduction to Bioinformatics**:
    - "*De-novo motif prediction in NF-KB pathways*" **–** Application of MEME tools for the prediction of de-novo motifs associated in NF-KB pathways and comparing it to TRANSFAC database to discover de-novo motifs in microarray data.
- **Introduction to Informatics**:
    - "Online patient data record"- Built a user interactive tool for doctors to retrieve online patient health data. Patient biographic data was stored in MySQL database. A user-interactive interface was built with PHP for accessing the patient data.

ACTIVITIES:

- **AAI-2010 (American association of Immunologists), Baltimore, MARYLAND.**
    - Poster : "DACS-DB: A dissemination and curation model to cure illness" in June 2010
- **IUPUI Research day, IUPUI, Indianapolis, INDIANA**
    - Poster : "DACS-DB: A dissemination and curation model to cure illness" in April 2010
- **ISBB-2006, 15-17 December, Bhubaneswar, ,INDIA**
    - Poster was awarded Best Poster in the conference titled "Sequence Analysis of H5N1 & H1N1 for Ascertaining Test System"
- **Member of Indiana Clinical and Translational Sciences Institute**