CYTOKINE POLYMORPHISM CATALOG (CytoCAT)
FOR
THE ANALYSIS OF PHENOTYPE ASSOCIATIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKÇE OĞUZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

JANUARY 2014

CYTOKINE POLYMORPHISM CATALOG (CytoCAT) for the Analysis of Phenotype Associations

Submitted by **Gökçe Oğuz** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics Program, Middle East Technical University** by,

Prof. Dr. Nazife Baykal ____________________________
Director, **Informatics Institute**

Assist.Prof. Dr. Yeşim Aydın Son ____________________________
Head of Department, **Health Informatics, METU**

Assist.Prof. Dr. Aybar Can Acar ____________________________
Supervisor, **Health Informatics, METU**

Assist.Prof. Dr. Yeşim Aydın Son ____________________________
Co-supervisor, **Health Informatics, METU**

Examining Committee Members:

Assoc. Prof. Dr. Tolga Can ____________________________
Computer Engineering, METU

Assist.Prof. Dr. Aybar Can Acar ____________________________
Supervisor, Health Informatics, METU

Assist.Prof. Dr. Yeşim Aydın Son ____________________________
Co-supervisor, Health Informatics, METU

Assoc. Prof. Dr. A. Elif Erson Benson ____________________________
Biological Sciences, METU

Asst. Prof. Dr. Bala Gür Dedeoğlu ____________________________
Biotechnology, Ankara University

Date: 27.01.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Gökçe Oğuz

Signature : _________________

ABSTRACT

CYTOKINE POLYMORPHISM CATALOG (CYTOCAT) FOR THE ANALYSIS OF PHENOTYPE ASSOCIATIONS

OĞUZ, GÖKÇE
M.Sc., Bioinformatics Program
Supervisor: Assist. Prof. Dr. Aybar Can Acar
Co-supervisor: Assist. Prof. Dr. Yeşim Aydın Son

January 2014, 84 pages

Currently, many studies focus on identifying disease related biological biomarkers for prediction of susceptibility, early detection, and prevention, in addition to developing new therapeutic approaches. In our study, we have investigated the single nucleotide polymorphisms (SNPs) of human cytokines and cytokine receptors, which play an important role in the immune system, as potential disease biomarkers and focused on phenotypes as they might give clue about disease symptoms. Our main aim was to create a catalog to be a guideline to early diagnosis, disease treatment, and drug discovery and design studies by storing general and specific associations between genes, SNPs and phenotypes. For that reason, firstly genetic variations on known human cytokines and cytokine receptors and then associations between these variations and phenotypes are identified. In particular, the data integration approaches were used to map single nucleotide polymorphisms (SNPs) on known cytokine and cytokine receptor genes and to extract SNPs associated to phenotypes from various biological databases. By congregating these data a new biological relational database was developed. A case study is done to further analyze and visualize the GWAS results for 3 different cancer types accessed through database of Genotypes and Phenotypes (dbGaP). This relational database enables one to search with different parameters and to analyze the associations from different aspects. As a result, a catalog of cytokine and cytokine receptor SNPs and their association with diseases is developed. This allows analysis of molecular and clinical research data from different perspectives, like identifying underlying etiology of diseases through associated polymorphisms, and SNPs common to cytokine-dependent diseases.

Keywords: Cytokine and Cytokine Receptor, Single Nucleotide Polymorphism, Database, Phenotype, Disease

ÖZ

FENOTİP İLİŞKİLERİ ANALİZİ İÇİN SİTOKİN POLİMORFİZM KATALOĞU(CytoCAT)

OĞUZ, GÖKÇE
Yüksek Lisans, Bioinformatics Program
Danışman: Yrd. Doç. Dr. Aybar Can Acar
Eş Danışman: Yrd. Doç. Dr. Yeşim Aydın Son

Ocak 2014, 84 sayfa

Son yapılan araştırmalar hastalığın tedavisinin yanı sıra önceden teşhis edebilmeye de odaklanmaktadır. Hastalık ön teşhisinde kullanılan çeşitli yöntemlerden biri de moleküler biyomarkerlar kullanarak hastalığa sebep olan genetik bozuklukları belirlemektedir. Bu çalışmada bağışıklık sisteminde çok önemli rol oynayan sitokinler ve sitokin reseptörleri genlerinde ortaya çıkan tek nükleotid polimorfizmleri(SNPler) potansiyel hastalık biyomarkerı olarak kullanıldı ve hastalık belirtilerine dair ipucu verebileceğinden dolayı fenotiplere odaklanıldı. Bu çalışmanın esas amacı erken teşhis, hastalık tedavisi ve ilaç keşfi ve tasarımı çalışmalarına gen, SNP ve fenotip arasındaki genel ve özel ilişkileri sunarak yardımcı olacak bir katalog geliştirmekti. Bundan dolayı öncelikle sitokin ve sitokin reseptörleri genlerinde bulunan SNPler ardından da bu SNPlerle fenotipler arasındaki ilişkiler tanımlandı. Bu çalışma için veri entegrasyonu yöntemleri kullanılarak, insan sitokinlerinin ve reseptörlerinin genlerinde meydana gelen tek nükleotid polimorfizmleri ve hastalıklara sebep olan tek nükleotid polimorfizmleri farklı biyolojik veritabanlarından edinildi ve bu bilgiler doğrultusunda yeni bir biyolojik veritabanı oluşturuldu. dbGaP sitesinden edinilen 3 farklı kanser tipiyle örnek çalışma yapıldı. Bu yeni veritabanı sayesinde kullanıcılar farklı parametrelerle araştırma ve farklı açılardan gen-hastalık ilişkileri analiz etme olanağı bulacaklar. Sonuç olarak en çok hastalığa sebep olan genler ve polimorfizmleri, farklı hastalıklardaki ortak mutasyonların bilgisini de içeren sitokin ve sitokin reseptörleri genlerindeki SNPlerin ve bunların hastalıklara ilişkininin kataloğu oluşturuldu.

Anahtar Kelimeler: Sitoin ve Sitokin Reseptörleri, Tek Nükleotid Polimorfizmi, Veritabanı, Fenotip, Hastalık

To my father Necmettin OĞUZ

and

My mother Figen OĞUZ

ACKNOWLEDGEMENTS

Special thanks to my supervisor Assist.Prof. Dr Aybar Can Acar and co-supervisor Assist.Prof. Dr. Yeşim Aydın Son for their encouragement, priceless guidance, continuous support and friendly attitude throughout my research.

I would like to thank Assoc. Prof. Dr. A. Elif Erson Benson, Assoc. Prof. Dr. Tolga Can, and Assist. Prof. Dr. Bala Gür Dedeoğlu for reviewing my work.

Special thanks go to my colleagues Adnan Öztürel and Fatih Ömrüuzun for their endless support, encouragement, and suggestions since the beginning of this research.

I would like to thank to Aykut Mert Yakut for his endless friendship and countless take outs even though world’s most expensive gasoline is used in Turkey.

I would like to especially thank to Okan Bilge Özdemir for answering all my never ending questions and being always there when I need courage, help and a shoulder to cry on.

I also thank to my friends Nurcan Alkış, Bilge Sürün, Kerem Eryılmaz, Elif Gök, Umut Çınar, Ersin Karaman, Gökçen Yılmaz, Banu Aysolmaz, Deniz İren, Özlem Özkan for their support during my study. I would like to thank to Informatics Institute members Sibel Gülnar, Hakan Güler and Yaşar Sayın for their support.

I would like to thank to Fulya Eren, Ceren Yıldırım, Selce Çakmacık, Buket Yiğit, Burcu Yaldız and my grandmother Müzeyyen Oğuz for their endless patience to meet again.

I also would like to thank Negin Bagherzadi for this last two months. She was a perfect lab friend, without her encouragements this thesis would not even be finished.

I would like to especially thank to Biber for her endless affection. Since we met, she never left me alone, even for a second.

Special thanks to my family for their endless support, love, trust, patience and encouragement during this study. They devoted their life on me.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

dbGaP: database Genotype and Phenotype

SNP: Single Nucleotide Polymorphisms

GWAS: Genome Wide Association Studies

ImmPort: The Immunology Database and Analysis Portal

MeSH: Medical Subject Headings

CHAPTER 1

Introduction

1.1. Motivation

The World Health Organization published a fact sheet in July 2013 showing that top nine of ten causes of deaths worldwide between 2000 and 2011 were related to some disease. Scientists are improving disease treatment methods to decrease death rates in direct proportion to developing technologies, but success rate of the treatments has not reached the rate of increase in cases. Likewise, today one of the main problems in medical genetics is to identify genetic variations that are responsible for diseases which are inherited from generation to generation. For that reason, more studies have started to focus on developing new methods to identify human gene variations, which resulted in agglomeration of new data of human single nucleotide polymorphisms (SNPs). As a result of the Human Genome Project, whole human genome sequence knowledge has been established. When all this information is combined, new ways of understanding the relationship between genotype and disease is possible. However, parts of this information is stored and analyzed in different databases and written in thousands of different articles. So in this research, we aimed to collect and integrate relevant data from different data sources and present them in a web based catalog.

Researchers interested in medical genetics usually focus on cytokines in their studies as they have a broad spectrum of function and effect. In addition to their vital role in both innate and adaptive immune systems, cytokines also affect non-immune cells which makes them suitable for use as biological response modifiers for clinical disease causing variation studies.

Although recent studies concentrate on cytokine and disease relationships, which enables to understand the underlying causes of diseases, they focus on specific cytokine types and/or specific diseases. As a result of this, the information about disease-cytokine relation increases, however revealed and analyzed in different articles. For that reason, to see the big picture, revealed results of these studies should be integrated and presented altogether.

Up to this time, for this purpose, two different research groups presented web-based solutions. First research group, Bidwell et al. focused on this problem in 1999 and supported their study with both in-vivo and in-vitro methods. However, the last update to the website was done in 2006 and now it is not available online. Second study was done by Bhushan, in 2011. DAC-DB is still online, however the website only stores disease and cytokine relationships.

Although cytokines are crucial molecules for the immune system, to function they need to bind to their receptors. Consequently, this relation makes cytokine receptors also crucial in immune system. Therefore, in this study one of our aims is to fill the knowledge gap of relation between cytokine receptors and diseases. Furthermore, we aim to find the symptoms associated with cytokine and cytokine receptor variations, for that reason, unlike other studies, in this study we focus on phenotypes, in addition to diseases.

Gene associated phenotype information will enable researchers to early diagnose patients and to start early treatments for people who might develop diseases in the future. Also, researchers who focus on drug design will benefit from this information in detection of drug target and chemical types to be used while designing the drugs.

What is more, in this era personalized medicine studies are increasing and scientists are focusing on developing drugs specific to individual patients rather than developing drugs for a symptomatic group. For that reason, by storing all cytokine and cytokine receptor genes, SNPs located on them and associated phenotypes, CytoCAT has the potential to be useful in personalized medicine studies.

1.2. Goal

The main objective of this thesis is to create a catalog which would help researchers during prediction of susceptibility and early detection for prevention, in addition to developing new therapeutic approaches for diseases.

For that reason, in this study single nucleotide polymorphisms are used as biomarkers to detect disease related phenotypes. As cytokines and cytokine receptors are one of the most crucial immunological molecules for both innate and adaptive immune systems, they are studied as target genes

Additionally, a web based catalog (CytoCAT) which shows association of phenotypes with single nucleotide polymorphisms located on the known cytokine and cytokine receptor genes is created. The PostgreSQL Database Management System is used to construct the database and in the web application results are currently presented in tabular view.

1.3. Outline of Thesis

This thesis is organized under five chapters including background and related works, datasets and methods, results and conclusion. In *Chapter 2,* background information about cytokine and cytokine receptor genes and single nucleotide polymorphisms is given. In addition to this, a literature survey on diseases related to cytokine and cytokine receptor and related studies is presented. In *Chapter 3,* data sources which are used to extract datasets and methods used to extract them are described. Later in Chapter 3, the database creation and entities, attributes and relations between them in the databases are explained in detail and presented as both ER diagram and relational table. In *Chapter 4*, statistics of database results and case study is presented. In *Chapter 5*, importance of cytokines and cytokine receptors, single nucleotide polymorphisms and CytoCAT is mentioned. In addition to these, significance of the records in CytoCAT is discussed. Lastly, in *Chapter 6,* achieved goals are mentioned and recommendations for further development of the catalog are given.

CHAPTER 2

Background and Related Works

2.1. Immune System

Every living thing has at least one mechanism to protect itself from pathogens such as chemicals, viruses and microorganisms. All together, these mechanisms form the "immune system". Immune system mechanisms are comprised of organs, cells and proteins and complexity of these mechanisms changes from organism to organism.

The human immune system is divided into two sub-types as the *innate* immune system and the *acquired* immune system. Individuals are born with all the components of their innate immune system, which defends the host from infections in first level. Natural barriers like skin and mucous membrane, chemicals such as cytokines, pattern recognition molecules are all components of innate immune system. The first aim of the system is to prevent pathogens from entering the host. If environmental agents get into the body, it recognizes and immediately responds. Although the system rapidly responds, it does not have any immunological memory, therefore it is not antigen specific.

When innate immune system is not able defend the host, the acquired immune system is triggered. It is comprised of cells and processes that are highly specialized to eradicate or prevent the growth of invaders. The main component of the system are lymphocytes in addition to antibodies and antigen recognition molecules such as B-cell and T-cell receptors. Although the acquired immune system responds slowly, it has immunological memory, therefore is antigen specific. After responding to an invading new pathogen, acquired immune system creates an immunological memory for that specific pathogen. This agent specific immunological memory ensures enhanced immune response to following invasion of the same pathogen.

2.2. Cytokine and Cytokine Receptors

2.2.1. Cytokines and Subtypes

First cytokine discovery was made in the late 1960s. In 1957 Isaacs A, Lindenmann J. published "*Virus Interference .I. The Interferon*" (Isaacs & Lindenmann, 1987). This was the first article mentioning interferons which became one of the milestones in cytokine studies. In 1972, Igal Gery and Byron Waksman published a paper which was the first time "lymphocyte activating factor" is mentioned (Gery & Waksman, 1972). Later in 1980, J.W. Mier and R.C. Gallo described T-cell growth factor for the first time (Mier & Gallo, 1980). Since then, researchers have continually attained knowledge about cytokine classification, function, mechanism, mutations and their disease relation.

Cytokines are described as messengers of both the innate and acquired immune systems. They are low molecular weight soluble proteins, peptides or glycoproteins (Costantini et al., 2009), which are secreted by immunocytes, epithelial and endothelial cells. They play an important role in the biological activities of the host's defense system against pathogens, functioning as mediators of inflammation, hematopoiesis, phagocytosis and apoptosis. (Borish & Steinke, 2003) In addition to this, they are involved in homeostasis, tissue repair, cell growth and development. (Mantovani et al., 2004)

"Cytokines may be classified on the basis of their cell of origin, their spectrum of activity, the category of activity they influence, the cells that are their targets, or on specific features of their ligand-receptor interaction." (Cohen & Cohen, 1996)

According to their targets and functions cytokines are divided into 5 sub-types as:

a. Interleukins
b. Interferons
c. Tumor Necrosis Factor Family
d. Transforming Growth Factor β Family
e. Chemokine

Interleukins are synthesized by various cells, but mainly by helper CD4 T cells. Activity of interleukins differs according to their members. In general, they stimulate the growth, differentiation and development of B cells, T cells and blood cells. (Coico, Sunshine, & Benjamini, 2003)

Interferons are secreted by T_H1 cells and are known with for general activity on viral infections by mediating the cellular response (Commins, Borish, & Steinke, 2010). They activate natural killer cells and macrophages. In addition to this, they upregulate major histocompatibility complex (MHC) class II expression which increases recognition of cells which are infected or converted to a tumor cell.

Tumor Necrosis Factor Family has two sub types. TNFα is secreted by macrophages and mast cells. It plays a major role in induction of fever and septic shock. On the other hand, TNFβ, also named as lymphotoxin, is secreted by T cells and it is involved in annihilation of target cells via cytotoxic CD8+ T cells. (Coico, Sunshine, & Benjamini, 2003)

Transforming Growth Factor β Family is secreted by lymphocytes, macrophages, platelets and mast cells. They play a role in immunoglobulin A production, growth of fibroblasts and wound healing. However, TGF β family members inhibits monocyte and T cell activation. (Coico, Sunshine, & Benjamini, 2003)

Chemokines are in 8-10 kilodaltons size and shares a structural characteristic to form tertiary structure which is 4 cysteine residues in conserved locations. Their main function is to direct immune cells by chemotaxis to the site of inflammation. (Commins et al., 2010)

2.2.1.1. Cytokine Mechanism

Cytokines conduct immune cell's interaction and communication via chemical signaling language, culminating in an immune response. They mobilize immune cells to the inflammation site where pathogens attack the host. In addition to this mobilization, they also trigger the activation of other immune cells. (S. Chen et al., 2007)

Chemical signaling of cytokines usually occurs locally and for a short time. Most of them act on their target cells in a *paracrine* or *autocrine* manner. When they act on the cell they are secreted, they are said to act in an autocrine manner. When we talk about paracrine manner, we mean that the secreted cytokines acts on the cells close to the cell which secretes the cytokine. Rarely, some secreted cytokines systemically travel through bloodstream and act on the cells at a distant site.
In the body, cells are never exposed to a single cytokine (Wood, 2006). Cytokines can affect multiple phenotypic traits by affecting the activity of different cell types differently, which indicates their pleiotropic property. Also, different cytokines may have the same function, which makes them functionally redundant. In addition to these properties, synergism and antagonism properties are also observed between cytokines. Transforming growth factor-β (TGF-β) increases the gene expression of type 1 collagen whereas tumor necrosis factor α (TNF-α) reduces the expression which is an example of antagonistic property of cytokines (Verrecchia & Mauviel, 2004). On the other hand, the production of interferon-γ (IFN-γ) by interleukin 12 (IL-12) and tumor necrosis factor-α (TNF-α) is an example of synergism (Ahlers, Belyakov, Matsui, & Berzofsky, 2001)

Cytokines are classified into "Pro-inflammatory" cytokines and "Anti-inflammatory" cytokines according to their response. Pro-inflammatory cytokines cause inflammation in the body. Most well-known pro-inflammatory cytokines are IL-1 and TNF which promote fever, inflammation, tissue reduction, shock and even death (Dinarello, 2000).On the other hand anti-inflammatory cytokines inhibit synthesis of pro-inflammatory agents. Most well-known anti-inflammatory cytokines are IL-4, IL-6, IL-10, IL-11 and IL-13. (Opal, 2000)

$CD4^+$ T cells differentiate into three subsets with respect to the cytokines they produce. T_H1 cells secrete cytokines which activate cell-mediated immune response like IL-2 and TNF-γ whereas T_H2 cells secrete antibody response affecting cytokines like IL-4 and IL-5. (Coico, Sunshine, & Benjamini, 2003) .Both T_H1 and T_H2 secreted cytokines interact with B cells and according to the subset of the T_H, B cells are induced to different Ig synthesis. By this way, cytokines involve in proliferation, differentiation and activation of immune response cells.

2.2.2. Cytokine Receptor and Subtypes

The receptors where cytokine ligands bind is defined as "cytokine receptors". Although cytokines play a vital role in immune system, without binding to a receptor, they are ineffective. In other words, cytokines can only affect target cells which express related receptors. Therefore, cytokine receptors are as crucial as cytokines in human immune system. Cytokine receptors are membrane proteins. Like cytokines, they might be soluble in addition to being membrane-bound. Many of them have subunit structures which are expressed according the affinity level. For example, when the affinity level is low or intermediate, IL_2 receptor is expressed with β and γ dimer whereas in high affinity level, the expressed subunit number increases to three (α, β and γ trimer). This shows that cytokine receptor expression is highly regulated. What is more, the varying expression levels and subunit chains guarantees that only target cells which are activated will form cytokine-cytokine receptor complexes. Another point is, some receptor subunits might be common and shared between receptors which leads cytokines to show synergetic and antagonistic properties. The main aim of receptor subunits is to transmit the signal as cytokines affect their targets by activating intracellular signaling.(Barrett, 1996)

Cytokine receptors are divided into 5 subtypes due to their structure and activity as:

a. Immunoglobulin superfamily receptors
b. Class I Cytokine receptor family
c. Class II Cytokine receptor family
d. TNF receptor superfamily
e. Chemokine receptor family

Immunoglobulin Superfamily Receptors have an immunoglobulin domain or at least an immunoglobulin like domain.

Class I Cytokine Receptor Family receptors are composed of cytokine recognition site and signal transduction site. These receptors are transmembrane proteins and shares a common amino acid motif: WSXWS. It is also defined as "Hematopoetin Receptor Family".

Class II Cytokine Receptor Family, similar to the class I cytokine receptor family, has two types of polypeptide chains. However, they do not contain WSXWS amino acid motif. As most of their ligands are interferon, theses receptors are also defined as "interferon receptor family".

TNF Receptor Superfamily has three divergent sub-types based on the motif in their cytoplasmic tail as death receptor, decoy receptor and activating receptor. (Coico, Sunshine, & Benjamini, 2003). Most of their ligands belong to TNF family. When they form receptor-ligand complex together, they dominate activated immune cell's life and death.

Chemokine Receptor Family whose ligands are chemokines, has a unique helical shape where 3 transmembrane domains are extracellular and 3 other transmembrane domains are in cytoplasmic structure. Also g-protein is used for signal transduction which makes cytokine receptors a member of G-protein coupled receptors.

2.3. Human Genome

The complete set of *Homo sapiens* genetic information is termed "Human genome". This genetic information is coded in cell nuclei on 23 chromosome pairs as DNA sequence. 22 chromosome pair contains the autosomal information where the last chromosome pair is allosome determining the sex.

Deoxyribonucleic acid (DNA) is built by repeating blocks of nucleotides. A nucleotide is composed of deoxyribose sugar, a phosphate group and a nucleobase. These nucleobases, Adenine

(A), Thymine (T), Guanine (G), Cytosine(C), are group as purines (A, G) or pyriminds (T, C) according to their biochemical structure. Two long biopolymers of DNA is formed by repetition of nucleotide units. According to James Watson and Francis Crick's study in 1953, (Watson & Crick, 1953) these two long polymers are shaped as double stranded helix. Ribose sugar and phosphate group forms the backbone of DNA while nucleobase of one strand binds with hydrogen bond to the nucleobase on the opposite strand pairing as one pyrimidine to one purine. Therefore, A binds T with 2 hydrogen bonds whereas G binds to C with 3 hydrogen bonds.

Every ordered DNA sequence which is found on a locatable region of a particular chromosome that encodes for a functional product is defined as a *gene* and they are the functional and physical units of heredity. Latest studies shows that these functional products coded by genes might be a subunit or a precursor of a protein, or an RNA molecule with important biological functions.

National Research Council proposed The Human Genome Project for the first time in 1988. The human genome project had 8 main aim (Pevsner, 2009)as:

1. Sequence the whole human DNA and map genes,
2. Identify the variations,
3. Improve sequencing technologies,
4. Progress in comparative genomics with model organisms,
5. Enhance functional genomic technologies
6. Build databases to store information,
7. Develop analyzing tools,
8. Explore and examine the ethical, legal and social issues

In 2001, International Human Genome Sequencing Consortium (IHGSC) published the draft version of the human genome sequenced and analyzed. Later, in 2004 the IHGSC reported that euchromatic sequence of human genome was finished. Main conclusions driven from the project was that a haploid human genome contains 3 billion DNA base pairs. What is more important, it was found that only 20,000-25,000 genes in other words only 2% of the human genome is protein coding. (Pevsner, 2009) The rest, 98% of the human genome is comprised of introns, non-coding RNAs, regulatory DNA sequences and repetitive DNA elements.

2.4. Single Nucleotide Polymorphisms

2.4.1. SNP

It is known that 90% of the genetic variations in human genome is single nucleotide polymorphism (SNP) which is defined as variation of one nucleotide in DNA sequence. These most common genetic variations take place in every 300 base revealing approximately 10 million SNPs is present in the human genome. If least frequent allele of a point mutation is present in at least 1% of the population, it is classified as Single Nucleotide Polymorphism.

Although most SNPs are bi-allelic, they might also be tri-allelic or tetra-allelic. As shown in the Figure 1, if the locus has only 2 varying nucleotides it is called bi-allelic, tri-allelic for 3 varying nucleotides and tetra-allelic for all 4 nucleotides. These variance in alleles affects utility of SNPs as markers.

SNP mechanism occurs in two ways as transitions or transversions. Transitions occur during the exchange of purines to other purines or pyrimidines to other pyrimidines whereas transversion occurs as pyrimidines exchange with purines. Although transversion has higher probability to happen, nature biases towards transitions (Vignal, Milan, SanCristobal, & Eggen, 2002).

Individual 1:…GAT**C**CGCT…CCTCC**C**AAAGTGC…GCATG**A**GCC…

Individual 2:…GAT**C**CGCT…CCTCC**T**AAAGTGC…GCATG**G**GCC…

Individual 3:…GAT**T**CGCT…CCTCC**A**AAAGTGC…GCATG**T**GCC…

Individual 4:…GAT**T**CGCT…CCTCC**G**AAAGTGC…GCATG**A**GCC…

Figure 1: Short DNA segment of same region from 4 random individual in a population. Showing bi-allelic(C, T), tri-allelic (C, A, T) and tetra-allelic (A, C, T, G) single nucleotide polymorphisms.

A SNP might be found in both coding and non-coding regions of DNA sequences in addition to intergenic regions. If a SNP occurs on regulatory region of non-coding/intergenic regions of DNA like promoter region, mRNA binding sites and splicing sites, transcription and mRNA stability levels increases or decreases, microRNA affectivity changes (Zienolddiny & Skaug, 2012) and SNP can strongly affect the phenotype (J. Wu & Jiang, 2013).

Coding region SNPs are classified to sub-types based on their effect on gene products. Synonymous SNPs are the single base changes which do not change the amino acid sequence therefore the product by means of degeneracy of genetic code. Non synonymous SNPs are the single base changes which results as a different amino acid sequence which may lead function and structure difference in proteins (Zienolddiny & Skaug, 2012). The non-synonymous SNPs also divides into two as missense SNPs and nonsense SNPs. Nonsense SNPs changes gene sequence which leads to an early stop codon. Codons that should be translated to synthesize the protein are not translated because of the early stop codon resulting incomplete and nonfunctional polypeptide sequences. Missense SNPs changes gene sequences causing a different amino acid codon during the transcription which causes translation of a different protein product. The examples of diseases caused by nonsense and missense SNPs might be exampled as beta zero thalassemia (Chang & Kan, 1979) and sickle cell anemia (Wang & Moult, 2001), respectively.

As it was mentioned before, one of the aim of Human Genome Project was to identify the gene variations. Thus more than 1.4 million SNP were mapped by Human Genome Project. Later, in 2007, The International Hapmap Consortium revealed 3.1 million SNPs on a haplotype map

(Pevsner, 2009). Currently dbSNP (the NCBI SNP database) contains 38,072,522 validated RefSNPs in build 137[1].

2.4.2. SNP as biomarker and comparison (Bower et al., 2013) to other biomarkers

A measurable characteristic which indicates some biological features is defined as a *biomarker*. These molecular markers are usually used to analyze biological processes, define disease causing agents or understand how the body reacts to drugs.

One of the most used biomarkers is variation of genome. In previous studies (Vignal et al., 2002), DNA variations were classified into three types based on the information they provide at a single locus:

- (a) The bi-allelic dominant
  - (i) Random amplification of polymorphic DNAs(RAPDs)
  - (ii) Amplified Fragment Length Polymorphisms(AFLPs)
- (b) The bi-allelic co-dominant
  - (i) Restriction Fragment Length Polymorphisms(RFLPs)
  - (ii) Single stranded Conformation Polymorphisms(SSCPs)
- (c) The multiple allelic co-dominant
  - (i) Microsatellites

Until SNPs were used as a biomarker, microsatellites were the dominant biomarkers, which are short (2-6 base pair) repeating sequences of DNA. (Turnpenny & Ellard, 2011). 3 to 100 of these repeats are observed on DNA, which leads to amino acid repetitions in 20% to 40% of protein sequences in mammals (Marcotte, Pellegrini, Yeates, & Eisenberg, 1999). When comparison between SNPs and microsatellites are made, differences are easily seen. Individual microsatellites vary more than individual SNPs. As a result, genotyping errors are more easily discovered when microsatellites are used as biomarkers. However, studies focus on SNPs more which makes SNPs more common than microsatellites. Therefore, SNP maps contains more information than microsatellite maps as biomarkers (Daw, Heath, & Lu, 2005).

[1] December 2013, http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=137

2.5. Diseases

2.5.1. Cytokine and Cytokine Receptor Related Diseases

"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." (Bhushan, 2011)

Interleukin 22 (IL-22) which is secreted by T helper 17 cells promotes protective and inflammatory effect in inflammatory bowel disease (IBD) by STAT3 signaling. Another diseases they are effective are ulcerative colitis (UC) and colon cancer (CC). Like in IBD, they activate the STAT3 signaling which inhibits apoptosis, promotes metastasis and causes tumor growth. (Jiang et al., 2013)

In 2008, Steinman studied cytokines role in three major human brain disorders, fever, multiple sclerosis (MS) and Alzheimer diseases (AD). In the study, it is revealed that IL-1, IL-6 and TNF-α alarm the hypothalamus for agents and indirectly trigger the febrile response result from induction of PGE_2 secretion. Another result from the study is, in the white matter of brain, an inflammatory response is triggered by IL-6, IL-17, TGF-β, IFN-γ which causes development of the characteristic pathology of MS, the plaques in the white matter. In addition to this, the most well-known treatment of MS is IFN-β which constricts the proinflammatory cytokines action. Although cytokines associated with AD are still unknown, in brain autopsy of AD patients, IL-1β, TNF-α and TGF-β are observed. In respect to their findings, Steinman hypothesized that blocking TGF-β receptors might give successful treatment results for both AD and MS (Steinman, 2008).

In 2003, van der Pouw Kraan et al. studied expression profiles of immunogenes via synovial fluid from early onset Rheumatoid arthritis (RA) patients and osteoarthritis (OA) patients as control. The results showed that IL-2, IL-4, IL-13, IL-17, IL-1, IL-15, EGF, bFGF are upregulated in RA patients. (van der Pouw Kraan et al., 2003) Likewise, in 2008, Su studied same data but with cDNA microarrays. The study results also showed that IL-15 upregulated in addition to T cell receptor β and γ chain, IL-6Rα and IL-6Rβ. (Su, 2008)

In 2002, Lock et al. studied MS gene profiles using brain biopsies of MS patients and controls at oligonucleotide microarrays. Results presented that IL-1R, IL-8R2, IL-11α, TNFR are all upregulated. Other studies found upregulation of IL-1R2, IL-15, IL-1RN, IL-1RA, CXCL2 and downregulation of BCL2, EGF, IL-8 and IL-17R as well as supported the findings of Lock et. al study. (Satoh et al., 2005; Sellebjerg et al., 2008; Stürzebecher et al., 2003)

Several studies compared the expression profiles of peripheral blood cells of Systemic Lupus Erythematosus (SLE) patients and controls using oligonucleotide microarrays. These studies unveiled that TNFSF10, TNFSF10C, TNFSF10D, IL-1α, IL-1R2, IL-1RAP, IL-8, CXCR1, CXCR1, IFN-ω are upregulated whereas IL-16 and CCR7 are downregulated. (Bennett et al., 2003; Han et al., 2003; Rus et al., 2002)

In 2001, Bowcock et al. studied Psoriasis which is a chronic inflammatory disease. In their study, by using oligonucleotide microarrays expression profile of normal, uninvolved and involved skin samples are observed. Only upregulated cytokine gene was IL-8 in this study (Bowcock et al., 2001), which was later, in 2003, supported by Zhou et al. study.(Zhou et al., 2003)

One of the autoimmune disease is Systemic sclerosis (SSc) which generally involves in internal organs like lung, heart and liver, but characterized by progressive sclerosis of the skin.(Kunz & Ibrahim, 2009) Different research groups studied gene expressions of this disease with different datasets such as fibroblast cultures, skin samples. In Withfield et al. study, CCR1, FGFR1, CTGF, FGF7 and SCYA19 expressed in high levels.(Whitfield et al., 2003) After two years, in 2005, Tan et al. analyzed same subject and results involved increase in expression level of PDGFC, FGFRL1 and decrease in VEGFB expression level. (Tan et al., 2005) In 2008, Milano et al. compared expression levels using samples extracted from different subtypes of SSc patients. The results demonstrated increase in FGF5, TNFRSF12A expression level and decrease in IL-15, CXCL5 expression level. (Milano et al., 2008).

2.6. Related Works

2.6.1. Disease Associated Cytokine SNP database (DACS-DB)

Disease Associated Cytokine SNP database was developed in 2011 by Sushant Bhushan. The database is freely available in http://www.iupui.edu/~cytosnp/. It is developed to guide biologists, immunologists and biomedical researchers to fill the knowledge gap of disease related cytokine SNPs. In the study, Bhushan collected, curated and annotated cytokine SNP data. As a result of final curation which was in October 2011, the database stores 456 cytokine genes, 63356 SNPS, 853 diseases and 4399 disease associated publication. In addition to cytokine-SNP-disease relation, DACS-DB presents heterozygosity allele frequency and function class of SNPs. For “disease association potential (DAP)”, a support vector machine (SVM) was applied. The SVM classified SNPs into two group as “disease” and “non-disease” with 74% accuracy. (Bhushan, 2011)

2.6.2. Cytokine Gene Polymorphism in Human Disease-Online Databases

In 1999, Bidwell et al. developed Cytokine Gene Polymorphism in Human Disease: Online Databases but currently it is not available online at http://www.pam.bris.ac.uk/services/GAI/cytokine4.htm. In this study, they focused on the allelic polymorphisms of cytokine genes and their disease association both in vivo and in vitro studies. In in vitro studies, Bidwell group examined the correlation between expression level of transcripts of cytokines and allelic polymorphisms on genes. In in vivo studies, the group tried to find the association between cytokine polymorphisms and diseases by extracting clinical outcome of related cytokine genotypes.(Bidwell et al., 1999) Following years till 2006, they published three supplementary papers informing that they updated the database with new cytokine polymorphisms associated to diseases.

The database was presenting the results of both in vitro expression studies and the in vivo disease association studies as two tables. In in vitro expression studies table gene symbol, polymorphism and allele (or haplotype), expression changes and reference was presented to the user. The expression attribute could get increased, decreased or not changed values only. In in vivo disease association studies table cytokine gene symbol and polymorphism, associated disease, significance of association (yes for significant, no for not significant) and reference were presented to the user.

2.7. Data Sources Used in CytoCAT

2.7.1. ImmPort

"The Immunology Database and Analysis Portal" (ImmPort) is developed by the Northrop Grumman Information Technology Health Solution team under the Bioinformatics Integration Support Contract (BISC) Phase II which is sponsored by the National Institutes of Health (NIH), National Institute of Allergy and Infectious Diseases (NIAID) and Division of Allergy, Immunology and Transplantation (DAIT) (The ImmPort, 2009) .

ImmPort is a portal where various biological information is produced, stored and analyzed. The scientific data stored in the portal is produced by researchers who are supported by NIAID/DAIT. ImmPort central data repository is created by integrating experimental data from NIAID/DAIT funded researches and public databases containing proteomic, genomic and immunological data. As mentioned before, the ImmPort has analyzing tools such as Flow Cytometry Analysis (FLOCK) and Major Histocompatibility Complex (MHC) Validation and Analyses. In addition to these, the system provides immunology focused ontology including disease ontology, gene ontology and cell type ontology. (The ImmPort, 2009)

Full access to using the analysis tools and the stored research data is only given to the DAIT funded research group members, government employees who are authorized, qualified researchers and NIH employees. However, guest users may access to the reference data which includes information about genes, SNPs, pathways, protein networks, MHC alleles and ImmPort gene lists. Also the system gives its users to do advanced search with multiple searching parameters.

2.7.2. Ensembl/BioMart

The Human Genome Project which is accepted as a milestone for life science researchers revealed three billion base pairs of manually annotated human genome sequence arousing the problem how researchers will access to the results. In 1999, before Human genome project ended, The Ensembl Project started as a collaborative project between the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) and Wellcome Trust Sanger Institute. The aim of this project was to solve this problem by developing a software which automatically annotates, integrates and presents the human genome data where researchers could easily access to the latest version. At the beginning Ensembl mainly focused on human genome, but as years go by they broadened their focus. Right now, after fifteen years, the October 2012 release covers 58 chordate and 3 non-chordate model organisms (Flicek et al., 2013) including *Danio reriro, Mus musculus, and Caenorhabditis elegance*. Ensembl Genome Browser is a bioinformatics framework where users get information about gene location, intronic and exonic regions of the gene sequence, how many transcripts that gene has, base pair length of transcripts and amino acid pair length of the corresponding protein. Furthermore, the browser provides comparative genomics including orthologous, paralogous information of the gene. In addition to these, users could find variation and regulation information supported by graphical views.

Although Ensembl interacts with different data sources, it uses its own identifiers. The unique and stable Ensembl identifiers are a set of numbers beginning with “ENSG” as a gene identifier whereas translations, transcripts, exons are identified with numbers beginning with ENSP, ENST, ENSE respectively (Hubbard et al., 2002). Query results are represented with these identifiers. However, external database links are also provided through result page.

The results of a biological experiment is not meaningful usually with bare eyes. Therefore scientists need to process and analyze the data by doing queries in biological databases. However, generally they need to do complex queries in multiple databases as each database stores different information. Ensembl team developed BioMart to address this problem. BioMart is a bioinformatics framework that enables users to arrange sub-databases from one or more relational databases (Baran, Gerner, Haeussler, Nenadic, & Bergman, 2011) using Ensembl data warehouse. The results are accessible through API for java, REST, SOAP, SPARQL in addition to the web interface (Guberman et al., 2011).

All BioMart views are standard. Firstly to create and build a sub-database, users need to choose a database. Ensembl BioMart presents five databases which are “Ensembl Genes 74”, “Ensembl Variation 74”, “Ensembl Regulation 74”, PRIDE (EBI UK) and “Vega 54”. When the database is selected, related datasets are listed. After selecting the relevant dataset, users narrow down their query preferring which filters they want to use. Finally attribute options which indicates the column names that will be presented in the output results are chosen. The opportunity of intersecting to dataset is also given to the users.

2.7.3. GWAS Central

Genome Wide Association Study (GWAS) aims to find if a variation is associated with a trait by analyzing common genetic variations in different individual genomes. These examined traits might be a phenotypic characteristic but generally studies focus on major diseases.

GWAS Central (www.gwascentral.org) is developed to visualize and access GWAS data. It is an openly accessible collection of summary level genetic association data and with its advanced tools the database allows its users to compare and discover relevant data sets from different aspects like genotypic, phenotypic or traits perspective.

GWAS central extracts data from various databases. These data resources are listed as (GWAS Central, 2013):

- National Human Genome Research Institute (NHGRI) GWAS Catalog
- Open Access Database Of Genome-Wide Association Results
- Japanese GWASdb
- Database of Genotypes and Phenotypes (dbGaP)
- Welcome Trust Case Control Consortium(WTCCC)
- Broad

- Cancer Genetic Markers of Susceptibility (CGEMS) Project
- 1958 British Birth Cohort
- Magic Consortium
- Spiro Consortium
- Giant Consortium

All these collected data sets are in different formats and details. Therefore, GWAS Central integrates them into a flexible and coherent data model. All the congregated data curated by removing duplicate markers, merging numerous data sets for discrete studies and reviewing the genetic markers for valid dbSNP rs identifiers. Furthermore, every study is evaluated for its scope of phenotype content. They are represented with a standardized ontology terminology to ensure that all phenotypes are standardized across all studies and to enable meaningful cross-study searches. Thus, phenotypes are mapped to MeSH (The National Library of Medicine-Medical Subject Headings) and HPO (Human Phenotype Ontology). (Beck, Hastings, Gollapudi, Free, & Brookes, 2013)

The GWAS Central users might do their queries in four different categories, genotype, phenotype, study list and markers. Except genotype every category has their own stable unique identifiers. Phenotype, Study and marker identifiers prefixed by ‘HGVPM’, ‘HGVST’ and ‘HGVM’, respectively.

Phenotype searches are done by using MeSH/HPO terms or by browsing phenotype trees. MeSH Phenotype tree has 11 main headings including ‘diseases’ and ‘chemicals and drugs’ whereas HPO phenotype tree has 2 main headings as ‘onset and clinical course’ and ‘phenotypic abnormality’. When a phenotype search is done by using MeSH/HPO term, users can limit their searches by changing p-value threshold and number of p-values in associated studies. As a result of this queries, information of the phenotype and its identifier, the related study name, its identifier and related study data, the phenotype ontology annotation and total p-values in that study are presented. Results can be ordered by identifier, phenotype or study name and related data can retrieve p-values or genes for those phenotypes.

Genotype search can be based on genomic location or HGNC gene symbol. The search can be limited by changing p-value threshold, 3’ flank and 5’ flank. As a result of this queries, information of the phenotype tested in study and its identifier, the related study name, its identifier and related study data, the phenotype ontology annotation and total p-values in that study are presented. Results can be ordered by phenotype, identifier and study name. The related data can retrieve genes, all p-values or p values in gene/region for those phenotypes.

Marker search can be based on dbSNP rs identifier or GWAS Central identifier. The search can be limited by changing p-value threshold, and by determining the number of markers per study. As a result of this queries, information of GWAS Central marker identifier, accession number, reference sequence coordinates, variation type, status, the sequence, marker related data link and links of other sites like OMIM, SNPedia and dbSNP are presented. The results can be ordered only by GWAS Central marker identifier. The marker related data link leads to another page which gives detailed information of that specific marker such as the study name and its GWAS Central identifier, the dataset used in that study, phenotype, its effect size, p-value, -log p-value and a related data link.

Study list searches are keyword oriented. These searches are done by study ID, authors, PubMed identifier and other text contained in the study titles. The search can be limited by changing p-value threshold and by determining the number of markers per study. The results return with information of GWAS Central study identifier, name of the study, associated phenotype(s), total p-values, related citations and related data links. The results can be ordered by of GWAS Central study identifier, name, number of markers and date created. Related data can retrieve p-values or genes for those studies. The marker related data link leads to another page which gives detailed information of that specific marker such as the study name and its GWAS Central identifier, the dataset used in that study, phenotype, its effect size, p-value, -log p-value and a related data link.

The summary of results can be exported in various formats. These file formats are listed as Excel spreadsheet, text file (comma separated, tab separated or space separated), RSS or Atom news feed and Json file format.

The latest version (11th version) of GWAS Central is released in September 5th 2013. The significant content of latest release comprise 1,605studies with 67,723,637 p-value and 2,935,163 unique dbSNP marker. With 67 million p-value for over 1600 studies, GWAS central is the world's largest publicly available collection of summary level GWAS information (GWAS Central, 2013);(Beck et al., 2013).

2.7.4. Medical Subject Headings (MeSH)

National Library of Medicine-Medical Subject Headings is a semantically controlled vocabulary. MeSH 2014 tree structure has 16 main categories in tree ontology as following: (National Library of Medicine, 2014):

1. Anatomy[A]
2. Organism[B]
3. Diseases[C]
4. Chemicals and Drugs[D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology[F]
7. Phenomena and Processes[G]
8. Disciplines and Occupations [H]
9. Anthropology, Education, Sociology and Social Phenomena[I]
10. Technology, Industry , Agriculture[J]
11. Humanities [H]
12. Information Science[L]
13. Named Groups[M]
14. Health Care[N]
15. Publication Characteristics[V]
16. Geographicals [Z]

CHAPTER 3

Material and Methods

3.1. Overview

In this chapter, sources of the data used in the study and their extraction steps are explained. The integration of this data is presented as a workflow. In addition to this, the database creation process is discussed with detailed explanation of all entities, their attributes and relation between those entities. Also this detailed explanation is visualized with ER diagrams and relational tables.

3.2. Data

3.2.1. ImmPort

In this research, ImmPort is used to extract cytokine and cytokine receptor gene lists. Protocol of the gene lists is briefly a species limited Entrez query which is done by using specific keywords with 'current only' option and saving as an excel file. Then the list is manually curated and assigned to sub-lists. As a result, both cytokine and cytokine receptor genes are classified under five subtypes in the portal:

Cytokine sub- types:

a) Interleukins
b) Interferons
c) TNF Family Members
d) TGF-β Family Members
e) Chemokines

Cytokine Receptor sub-types:

a) Interleukin Receptors
b) Interferon Receptors
c) TNF Family Member Receptors
d) TGF-β Family Member Receptors
e) Chemokine Receptors

ImmPort offers options to view lists according to gene ontology, protein interactions or gene summary. In our study, we preferred ImmPort gene summary table which provides Entrez gene id, gene symbol, gene name, synonyms and chromosome number of the gene. Both general gene tables and its sub-types gene tables are extracted from the portal in excel files for both cytokine and cytokine receptor genes. The number of data extracted for each list is given in Table 1.

*Table 1: The number of data in related gene lists and their sub-types list a) For cytokine gene lists and their sub-types lists. b) For cytokine receptor gene lists and their sub-types lists. * The genes which are not classified to a subtype are listed under other genes list.*

a)

Name of Gene List	**Number of Data**
Interleukins	47
Interferons	17
TNF Family Members	12
TGF-β Family Members	33
Chemokines	102
Other Cytokine Genes*	245
Total Cytokine Count	**456**

b)

Name of Gene List	**Number of Data**
Interleukin Receptors	43
Interferon Receptors	3
TNF Family Member Receptors	19
TGF-β Family Member Receptors	12
Chemokine Receptors	53
Other Cytokine Receptor Genes*	178
Cytokine Receptor Count	**308**

3.2.2. Ensembl/BioMart

In our study, BioMart is used three times. First one was to convert Entrez gene identifiers to Ensembl gene identifiers. Second one was to extract general gene information and third, to get dbSNP rsids located on the cytokine and cytokine receptor genes. For all uses, Ensembl Genes 74 and Homo sapiens genes (GRCh37.p13) were selected as database and dataset, respectively.

While converting Entrez gene identifiers to Ensembl gene identifiers, queries are limited by uploading gene ID lists to filter- gene option. As attributes Ensembl gene id, Entrez gene id, associated gene name, chromosome name, gene start (bp), gene end (bp), strand, band, associated gene database, percentage of the GC content and source are selected. These operations are both done for cytokine and cytokine receptor gene lists. The queries can be seen in Appendix A.

To create a sub-database presenting cytokine and cytokine receptor genes and their associated single nucleotide polymorphisms, Ensembl gene identifiers which we imported in first step are used in gene filter option. In addition to this under the variation filter, germline variation source is marked as dbSNP. As attributes associated gene name, Ensembl gene id, reference id, chromosome name, chromosome location (bp), Variant Alleles, minor allele, minor allele frequency, variation source, Evidence status are all selected. These operations are both done for cytokine and cytokine receptor gene lists. The queries can be seen in Appendix B.

3.2.3. GWAS Central

In our study GWAS Central is used to retrieve association between phenotype and single nucleotide polymorphisms on known cytokine and cytokine gene receptors. We have approximately 1 million distinct SNP rs identifiers which means doing single query one by one would be time consuming and require too much efforts. Therefore, we needed to do batch queries. Although GWAS Central comprise BioMart based system, GWAS Mart, it did not solve our problems. Even though it is not written in any document, GWAS Mart can only do batch queries with maximum 2500 identifiers, which means if we used GWAS Mart we should be running same query with different sets of identifiers for 400 times. Therefore, we wrote a python script to extract all the associated phenotypes with given list of SNPs. This result file, in other words summary report, is extracted in Json file format, consisting information of dbSNP rsid, GWAS Central marker identifier, phenotype, GWAS Central phenotype identifier, log 10 p- value, GWAS Central study identifier, and GWAS Central experiment identifier.

3.2.4. Medical Subject Headings (MeSH)

Although GWAS Central uses MeSH terms, in the summary report neither MeSH identifier nor the ontology tree identifier is given for phenotypes. Thus, to be able to group/cluster related phenotypes, we needed these informations. For this reason, the mesh tree 2014 is downloaded from the U.S. National Library of Medicine- MeSH website (http://www.nlm.nih.gov/mesh/filelist.html) and a python code (Appendix C) is written to extract interested information as mentioned in previous section.

3.3. Integration of Data

3.3.1. Flowchart

The flowchart of integrated data is presented in Figure 2.

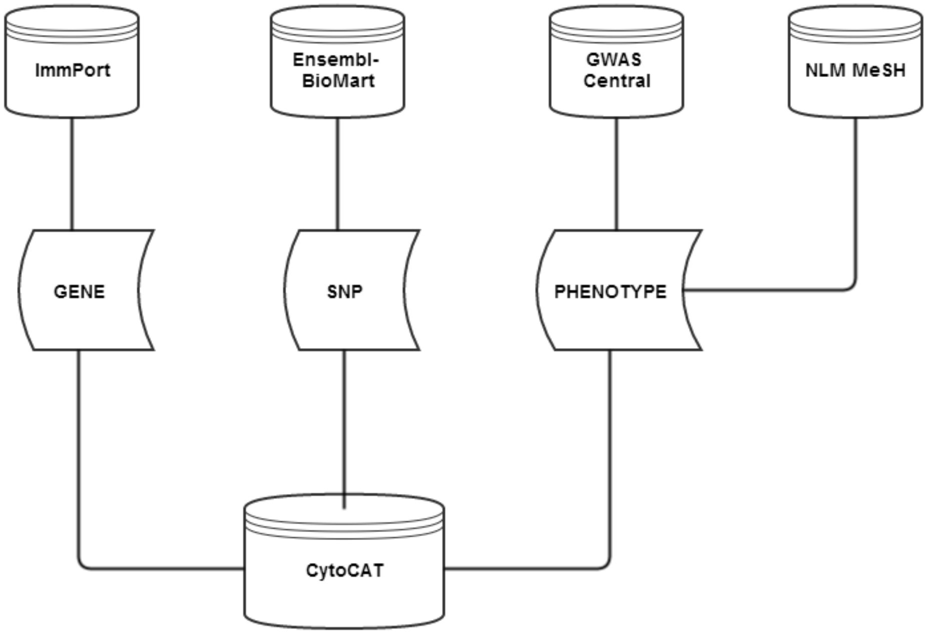


Figure 2: The flowchart of Data Integration. The data sources, which data sets are extracted and where they are imported can be seen in the flowchart.

3.4. Preprocessing Data

In this section, preprocessing of each data set is explained step by step. The gathered data from ImmPort, Ensembl-BioMart, GWAS Central and MeSH are examined, analyzed and processed before importing to the database. As the database accepts only comma separated values while importing data into tables, to convert data files into csv files, all the exported data sets are analyzed in excel. The data sets are manually curated and to ensure the consistency and accuracy, some attributes with wrong data are eliminated.

3.4.1. Preprocessing of ImmPort Data

The family member lists of cytokine and cytokine receptor genes extracted from ImmPort are directly used to extract information from Ensembl/BioMart.

3.4.2. Preprocessing of BioMart Gene Data

Firstly, the Ensembl gene identifier list that was converted from Entrez gene identifier is analyzed. The downloaded cytokine file had 523 rows. In the data, duplicating identifiers are observed. Ensembl and Entrez sites identifies and sometimes classifies genes differently, for that reason the genes assigned to identifiers might be different which causes these kind of duplications. For example, *CGB2* gene is identified as 'ENSG00000104818' in Ensembl database on the other hand Entrez identifies this gene with two identifiers, '114335' and '114336'. Likewise, *CALCA* gene is identified with '796' in Entrez database, however Ensembl identifies this gene with 'ENSG00000262933' and 'ENSG00000110680'. As in this study based on Ensembl gene identifiers, we removed the duplicated Ensembl IDs. As a result, we got 506 unique Ensembl Gene identifiers. Secondly, when we analyzed this 506 unique gene identifier, the Locus Reference Genomic database sourced genes has gene identifiers with LRG prefix. For same reason, we eliminated these 13 entries, after verifying that genes they assigned are also presented with Ensembl Gene identifier. Lastly the 38 entry with patch numbers rather than chromosome numbers are omitted. As a result, we got 457 cytokine gene entry ready to be imported to the database.

Same procedures are done to the cytokine receptor gene file where no duplicated identifier is observed. On the other hand, 18 genes with LRG identifier and 15 genes with different chromosome numbers are removed. As a result, 304 unique cytokine receptor gene information arranged to be imported to the database.

As mentioned before, the unique gene identifier lists were used in the second query at Ensembl/ BioMart to map single nucleotide polymorphisms to known cytokine and cytokine receptor genes. The downloaded cytokine file had 422344 rows. Duplications of gene identifiers with same dbSNP rsid are observed at 33 cytokine genes (Table 2). Different chromosome locations which are stated for the same variation on the same gene were causing those duplications. When the reason analyzed more deeply, it is that this is a well-known problem and explained in NCBI dbSNP website (dbSNP, 2008). The problem arouses from using various assemblies like the Reference assembly, the Celera assembly, the Venter Diploid assembly (HuRef) while mapping. In addition to this, the length of SNP's flanking sequence, the chromosome region SNP is mapping and variations in Flanking sequences of SNPs also affect this mapping on same chromosome to different locations. To be more specific, as flanking sequence gets longer, its chance to uniquely map to an assembly increases and if a SNP maps to a repetitive region, the possibility to mapping uniquely decreases.

Table 2: Gene symbols of Cytokine genes which have same rsid on different chromosome locations.

Cytokine Gene Symbol	**Cytokine Gene Symbol**	**Cytokine Gene Symbol**	**Cytokine Gene Symbol**
CGB1	CMTM8	IFNA7	CGB8
SEMA3A	CSHL1	IGF1	NRG1
NRG3	DEFA1B	IL16	PDGFRA
CGB2	FGF1	IL32	PDGFRL
CGB5	FGF12	IL33	RP11-204M4.2
CGB7	FGF13	IL37	CGB8
TG	GH1	LTBP1	SEMA5A
CGB	BDNF	SEMA3E	SEMA6B

Two sub-files, SNP-GENE association file and SNP file which contains general information about these SNPs, are generated from this cytokine file. SNP file which is comprised of reference id, variant alleles, minor allele, minor allele frequency, variation source, and evidence status, contains 408391 unique rs ID. On the other hand, SNP-GENE file which is comprised of associated gene name, Ensembl gene id, reference id, chromosome name and chromosome location (bp), contains 421118 unique association between cytokine gene and SNP.

Same procedures was done to downloaded cytokine receptor file. Likewise in cytokine file, duplications of gene identifiers with same dbSNP rsid are observed at 30 cytokine receptor genes as listed in Table 3. By using downloaded cytokine receptor file, SNP-GENE association file and SNP file which contains general information is generated with same attributes of corresponding cytokine files. As a result SNP file has 520317 unique rsid. In addition to this, SNP-GENE file has 525216 unique association between cytokine receptor gene and SNP at 525,421.

Table 3: Gene symbols of Cytokine receptor genes which have same rsid on different chromosome locations

Cytokine Receptor Gene Symbol	**Cytokine Receptor Gene Symbol**	**Cytokine Receptor Gene Symbol**
ACVR1C	IL3RA	PRLR
ADIPOR1	LEPR	PTH2R
BMPR1B	MCHR2	ROBO1
CALCRL	NR1H2	ROBO2
ESR1	NR1H4	RORA
ESRRG	NR1I2	RXFP1
FGFRL1	NRP1	TNFRSF14
FSHR	PLXNA2	TNFRSF8
HTR3B	PPARD	TSHR
IGF2R	PPARG	VIPR2

3.4.3. Preprocessing of GWAS Central Data

This data is procured from GWAS Central by both unique rsid lists of SNPs on known cytokine and cytokine receptor genes. The main purpose of this file is to indicate associations between SNPs and phenotypes, thus it is saved as PHNEOTYPE file. In this file, it is observed that 1836 of entries has no phenotype related information (phenotype, and phenotype id). Therefore these entries are deleted. When these 1836 entry is examined more deeply, it is revealed that they all belong to same study, HGVST890. Three experiment of this study which are HGVE1604, HGVE1605, and HGVE1606 are responsible for this missing phenotype information.1250 dbSNP rsid is affected from this missing data problem. However, before deleting rows with missing values, it is ensured that those affected SNPs are studied in different studies and experiments, therefore, there has been no data loss.

3.4.4. Preprocessing of MeSH Data

In this study, we are only interested in phenotypes, for that reason, every category analyzed deeply to ensure their terms are referencing to a phenotype. As a result, entries mapped to following categories are removed.

1. Phenomena and Processes[G]
2. Disciplines and Occupations [H]
3. Anthropology, Education, Sociology and Social Phenomena[I]
4. Technology, Industry , Agriculture[J]
5. Humanities [H]
6. Named Groups[M]
7. Publication Characteristics[V]
8. Geographicals [Z]

Mesh data has 55,611 records, however, in this research focused on only 88.5% (49,201 entries) of the data.

When table analyzed, mapping to some categories might be seen interesting. However, every entry is meaningful. For example, one of the MeSH term categorized under *Organism* title are all related to *'HIV-1 susceptibility'*. Another example is the MeSH terms categorized under *Analytical, Diagnostic and Therapeutic Techniques and Equipment* are '*Blood cell count'* and *'Respiratory Function Tests'.*

3.5. Database

In a database, entities are represented with tables whereas attributes are shown as columns of tables. The data which will be stored in the database are entered as rows of those entity tables. Each row, in other words tuple, must be unique. The most important constraints of an entity table are primary key and foreign key. Primary keys are the set of attributes which have unique values, therefore making the tuple unique. Foreign keys are used to provide data access between different tables. It can be defined as the attribute of a table uniquely defining primary key of another table.

3.5.1. ER Diagram

A conceptual data model is needed to be produced to reflect the structure of the data in the database. Therefore, an entity relationship diagram (ER Diagram) (P. P.-S. Chen, 1976) is generated to show the entity tables, their attributes and relation between these entity tables. The ER diagram of CytoCAT which is summarizing the relations between tables, giving information about the attributes and primary keys of entities is shown in Figure 3. What is more, which tables has one to one, one to many, many to many relations can be easily interpreted from the diagram.

3.5.2. Database creation

3.5.2.1. PostgreSQL

The database is PostgreSQL, an open source object relational database management system which is developed by PostgreSQL Global Development Group. It runs on all major operating systems, including Linux, UNIX and Windows. It has evolving since 1995. In this study, we used PostgreSQL 9.3.1 release. Psql, the primary command-line program and pgAdmin, the free and open source graphical user interface are used as database administration tool for PostgreSQL (Momjian, 2001).

3.5.2.2. CytoCAT Database Setup

The data gathered from ImmPort, BioMart and GWAS Central is stored in entities of PostgreSQL database. The data definition language (DDL) of CytoCAT entities presented in Appendix D and the relational data model of CytoCAT is shown in . CytoCAT consists of 9 tables, 5 entities and 4 connection tables. Functions and attributes of each table will be explained in details below:

GENE: This is the entity where all the information about interested genes is stored. As mentioned above the cytokine and cytokine receptor gene lists are extracted from ImmPort and more information about these gene are retrieved from Ensembl BioMart. This entity stores:

- Ensembl Gene identifier (id),
- Gene symbol (symbol),
- Chromosome number of gene located on (chr_no),
- Gene starting and ending base pair (gene_start, gene_end),
- Band and strand gene located on (band and strand),
- Percentage of the G-C base content (gc_content),
- The source database of related gene (gene_db) and
- The family identifier gene belongs to (familyid).

As every gene has a unique gene identifier, the primary key of this entity is ‘id’. ‘Familyid’ attribute is foreign key that refers to familyid attribute of Family entity.

SNP: This is the entity where all the information about interested single nucleotide polymorphisms is stored. As mentioned above the list of SNPs on known cytokine and cytokine receptor genes are extracted from Ensembl BioMart. This entity consists of SNP’s:

- dbSNP rs identifier(rsid),
- Varying alleles(var_allele),
- The least found allele(minor_allele),
- The frequency of the least found allele(minor_allele_freq),
- The list of sites/projects conforming this information (evidence status).

The primary key of SNP entity is the set of ‘rsid’ and ‘var_allele’.

GENE- SNP: This is rather a connection table than entity. It is formed to link the genes and SNPs related to this genes by pairing genes and SNPs. It has four attributes as:

- Gene_symbol,
- ensid,
- rsid,
- Chr_loc.

The chr_loc attribute stores the information of the chromosome location where SNP observed on the gene. Rsid attribute refers to the rsid in SNP entity and gene_symbol and ensid refers to the gene_symbol and ID in GENE entity respectively. Rsid, geneid and chr_loc all together forms the primary key of the table.

FAMILY: This entity stores the types of cytokine and cytokine receptors. The information stored in this entity is extracted from ImmPort. Family entity has three attributes as:

- Family identifier (familyid),
- The name of the family (family_name) and
- Parent of the family belongs to (parentid).

FamilyID is the auto incrementing primary key of this entity where each number assigned to a family. Parentid gets either ‘1’ or ‘2’. If the family belongs to a cytokine gene type, parenteid gets value ‘1’ whereas cytokine receptor families get ‘2’ as parentid.

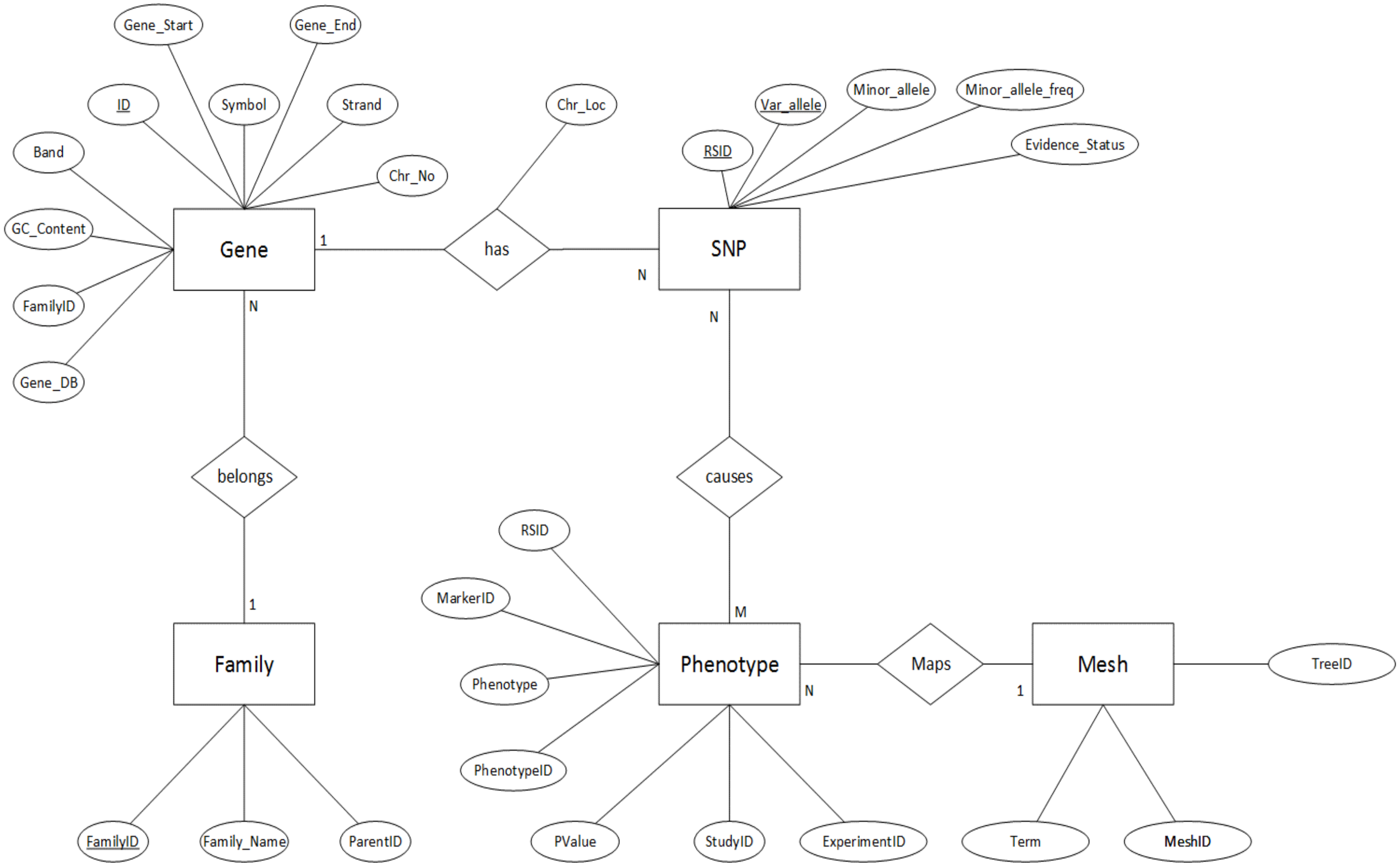


Figure 3: The ER Diagram of CytoCAT. The abstract version of CytoCAT which is summarizing the relations between tables, giving information about the attributes and primary keys of entities.

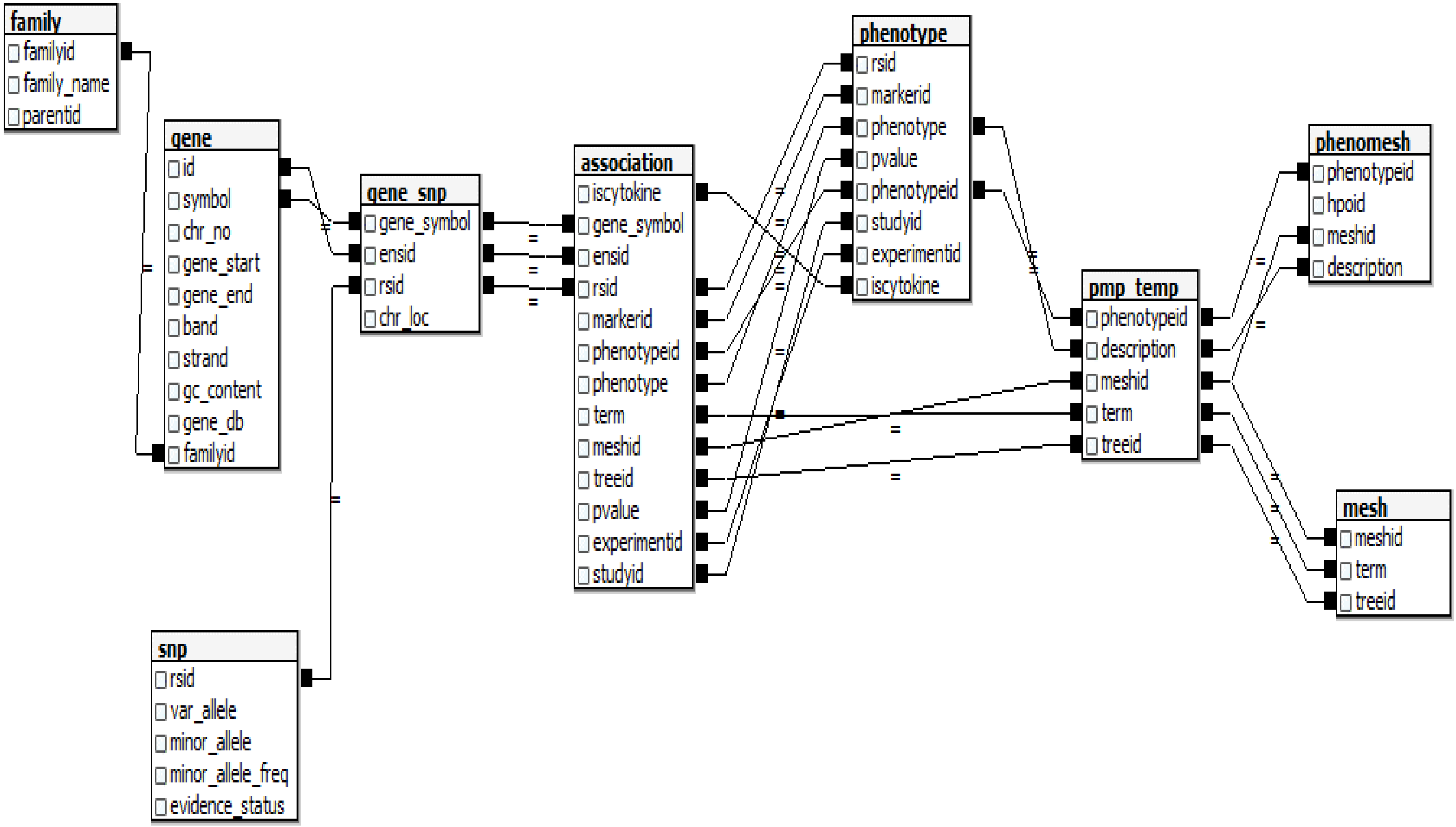


Figure 4: The Relational Data Model of CytoCAT. The final version of CytoCAT which is summarizing the relations between tables, giving information about the attributes and primary keys of entities

PHENOTYPE: This entity is created to store the data extracted from GWAS Central. The main purpose of this table is to show all phenotype-SNP associations in one table. It stores the phenotype information of cytokine and cytokine receptor related SNPs. Therefore this entity is the main table and it is linked to other master tables, SNP and GENE. This has 8 attributes:

- dbSNP identifier of variation(rsid),
- GWAS Central marker identifier of the variation (markerid),
- Variation associated phenotype(phenotype),
- - log p-value of the variation(p-value),
- GWAS Central phenotype identifier(phenotypeid),
- GWAS Central identifier of study which reveals the association between phenotype and SNP(studyid),
- GWAS Central identifier of experiments which reveals the association between phenotype and SNP(experimentid)
- Boolean value indicating if the data associated with a cytokine gene (isCytokine).

The rsid and markerid attributes are different identifiers of different databases but they both represent same single nucleotide polymorphism. The aim of the ‘pvalue’ is to show how significant the result of the experiment. ‘isCytokine’, as mentioned above, specifies if the data related to a cytokine or cytokine receptor gene. As it is a boolean attribute, it returns true when a cytokine gene-snp-phenotype association is represented and turns ‘False’ when the association is shown between cytokine receptor gene-SNP-phenotype. The table has two index constraints for phenotypeid (phenoidx) and markerid (markeridx).

MESH: This table is formed to understand and classify phenotypes more easily. The table has 3 attributes:

- Term
- Meshid
- Treeid

Term is the MeSH term of the phenotype. Meshid represents the MeSH identifier of the term. As MeSH is the semantically phenotype ontology, treeid indicates the location of the term on the ontology tree. Although every term has a unique meshid, they might have multiple treeids. Therefore the primary key of this table is ‘treeid’ attribute.

PHENOMESH: This table is a connection table between Phenotype entity and MeSH table. It has 4 attributes:

- GWAS Central phenotype identifier (phenotypeid)
- Human Phenotype Ontology identifier (hpoid)
- Medical Subject Headings identifier (meshid)
- Phenotype Descriptor(description)

phenotypeid is the primary key. The main purpose of this table is to map GWAS Central phenotype and phenotypeid to meshid. Later this table is used to catalog GWAS Central phenotypes and phenotypeids according to the specific MeSH tree nodes and branches.

PMP-TEMP: This table is a connection table between Phenotype, Mesh and PhenoMesh tables: It has 5 attributes as:

- MeSH term (Term)
- MeSH tree structure identifier(Treeid)
- GWAS Central phenotype identifier (phenotypeid)
- Medical Subject Headings identifier (meshid)
- Phenotype Descriptor(description)

Its primary key is formed by treeid and phenotypeid

ASSOCIATION: This entity is the main table. ıt stores information from every entity in CytoCAT. Its attributes can be listed as:

- Boolean value indicating if the data associated with a cytokine gene (isCytokine)
- Gene symbol (gene_symbol),
- Ensembl Gene identifier (ensid)
- dbSNP rs identifier(rsid),
- GWAS Central marker identifier of the variation (markerid)
- GWAS Central phenotype identifier (phenotypeid)
- Variation associated phenotype(phenotype)
- MeSH term (Term)
- Medical Subject Headings identifier (meshid)
- MeSH tree structure identifier(Treeid)
- p-value of the variation(p-value)
- GWAS Central identifier of study which reveals the association between phenotype and SNP(studyid)
- GWAS Central identifier of experiments which reveals the association between phenotype and SNP(experimentid)

CHAPTER 4

Results and Case Study

4.1. Database Result Statistics

4.1.1. Gene

In our database, we have 2 groups of genes: Cytokine genes and Cytokine Receptor genes. All in all, there are 765 related genes. 457 (60 %) of them belong are cytokines and 308 (40%) of them are cytokine receptors.

These genes are divided into subgroups based on the family they belong to. Each gene type has six families. Cytokine families, their count and percentage distribution can be seen in Figure 5. Likewise, cytokine receptor families, their count and percentage distribution can be seen in Figure 6.

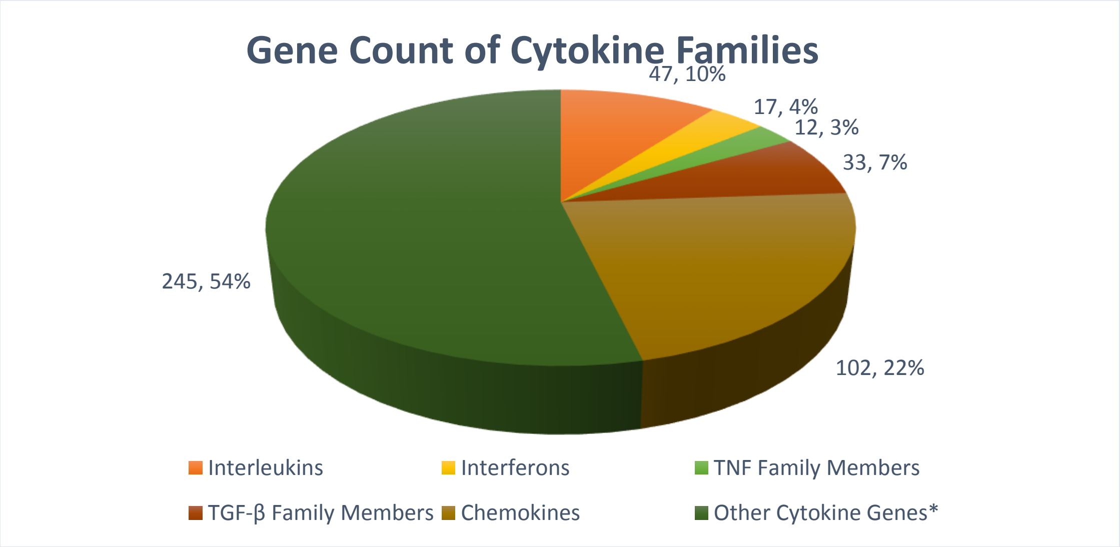


Figure 5: Pie chart of cytokine gene families, count and percentage distribution of genes belong to families.

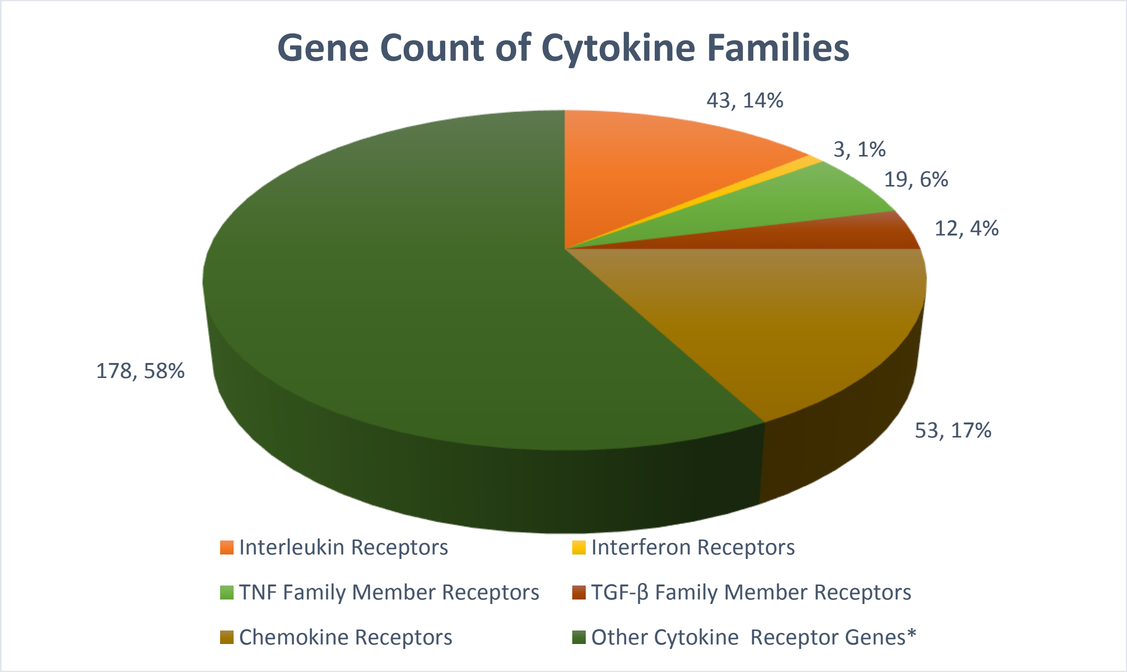


Figure 6: Pie chart of cytokine receptor gene families, count and percentage distribution of genes belong to families.

The source of 757 genes is HGNC symbol whereas two genes comes from Clone-based (Vega) and only one of them belongs to UniProtKB Gene. Genes are all distributed to 23 chromosome. Distribution of genes on chromosomes is represented in Figure 7. 377 of the genes are located on the +1 strand of genome whereas 384 of them are located on -1 strand of genome. These genes are dispersed on 173 distinct bands, changing from p11.2 to p36.33 and q11.1 to q43. Most of them intensified on "p13.3", “q12", “q13.3", "p21.3" and "q13.33". The percentage of GC base pair content has a very wide range changing from 29.02% to 73.35% in 672 unique value.

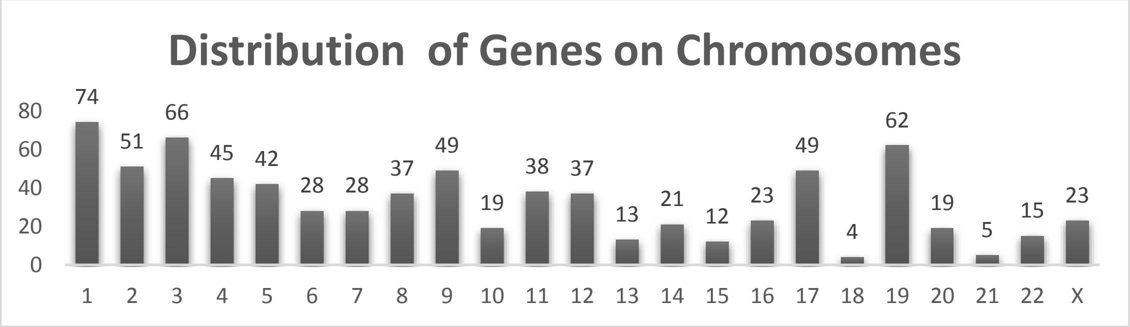


Figure 7: Distribution of genes on chromosomes stored in CytoCAT.

4.1.2. SNP

The SNP table is comprised of 926601 unique rsid and it preserves 12452 different variant alleles. There are 779 distinct minor alleles and their frequencies range between 0.0005 and 0.251374. The distribution of twelve most varying alleles is presented in Figure 8. When minor alleles are sorted by the number they have observed, thymine is in the first place with 188346 times. The second most observed minor allele is adenine with 186946 times which is followed by guanine as it has been observed 139506 times as a minor allele. Latter base is cytosine which has been observed 135021 times as a minor allele.

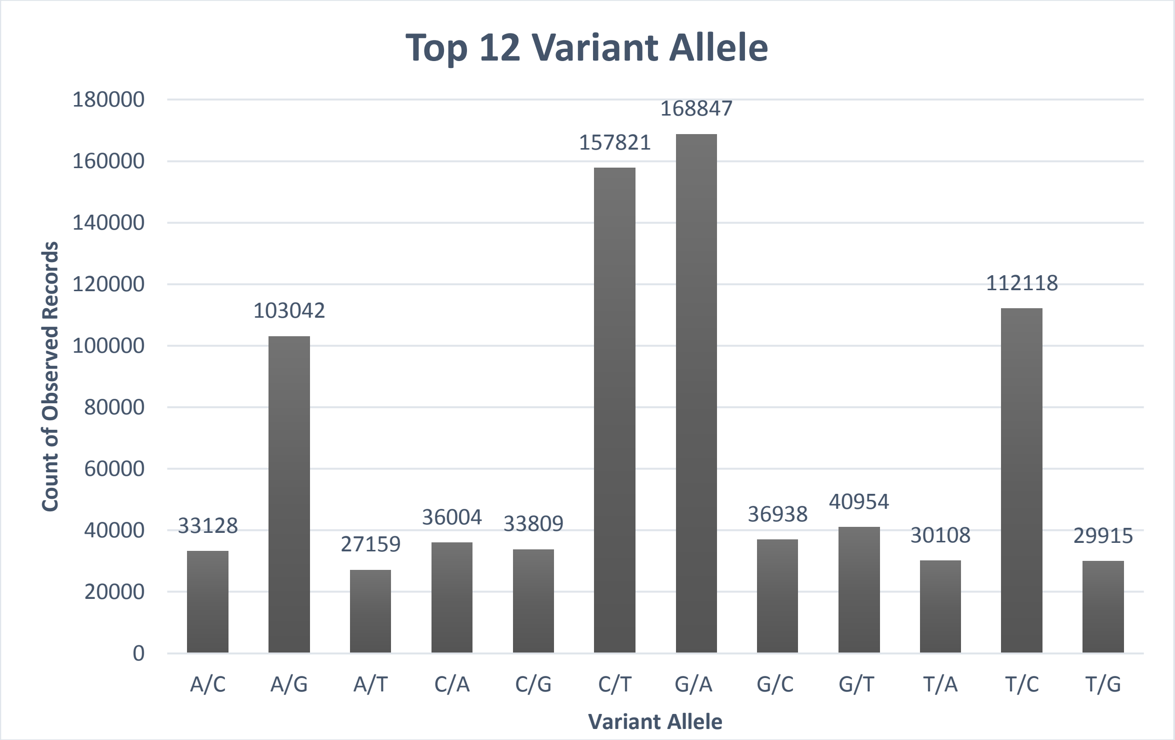


Figure 8: The most varying twelve base pairs found in CytoCAT. A stands for Adenine, C stands for Cytosine, G stands for Guanine and T stands for Timin.

As the main source all of this SNP information is dbSNP, the evidence status attribute values are extracted from dbSNP. dbSNP uses different sources to be more accurate which are Multiple observations, Frequency, HapMap, 1000 Genomes, Cited, Exome Variant Server (ESP). Generally these sources are not used singly, multiple sources are used as evidence, which can be observed in Table 4.

Table 4: The multiple sources used as an evidence for related SNPs and their counts. This evidence status are extracted from dbSNP and stored in SNP entity in CytoCAT.

Evidence Status	**Count**
1000 Genomes	138959
Multiple observations, Frequency, 1000 Genomes	120140
Multiple observations, Frequency, HapMap, 1000 Genomes	57533
Frequency, 1000 Genomes	53836
Multiple observations	48271
Frequency	41130
Multiple observations, Frequency	21638
Multiple observations, 1000 Genomes	20717
ESP	9248
Frequency, ESP	8386
Multiple observations, Frequency, 1000 Genomes, ESP	6537
Multiple observations, Frequency, ESP	4991
Multiple observations, Frequency, HapMap, 1000 Genomes, Cited	4812
Multiple observations, Frequency, HapMap, 1000 Genomes, ESP	1995
1000 Genomes, ESP	1139
Multiple observations, Frequency, HapMap	1109
Multiple observations, Frequency, 1000Genomes, Cited	830
Multiple observations, Frequency, HapMap, 1000 Genomes, Cited, ESP	792
Frequency, 1000 Genomes, ESP	473
Multiple observations, 1000 Genomes, ESP	362
Multiple observations, ESP	283
Multiple observations, Frequency, 1000 Genomes, Cited, ESP	253
Frequency, HapMap	209
Cited	147
Multiple observations, Cited	141
Multiple observations, Frequency, Cited	127
Multiple observations, Frequency, HapMap, ESP	90
Multiple observations, Frequency, HapMap, Cited	84
Multiple observations, Frequency, Cited, ESP	50
Multiple observations, 1000 Genomes, Cited	44
Frequency, Cited	43
Multiple observations, Frequency, HapMap, Cited, ESP	15
Cited, ESP	8
Multiple observations, Cited, ESP	6
Multiple observations, 1000 Genomes, Cited, ESP	3
Frequency, HapMap, ESP	2
Frequency, Cited, ESP	2
1000 Genomes, Cited	1
Frequency, 1000 Genomes, Cited	1

4.1.3. Gene_SNP

This table has 947,765 records. 422,344 of them represent the relation between cytokine genes and SNPs. The remaining 525,421 entries represent the relation between cytokine receptor genes and SNPs which is visualized in Figure 9.

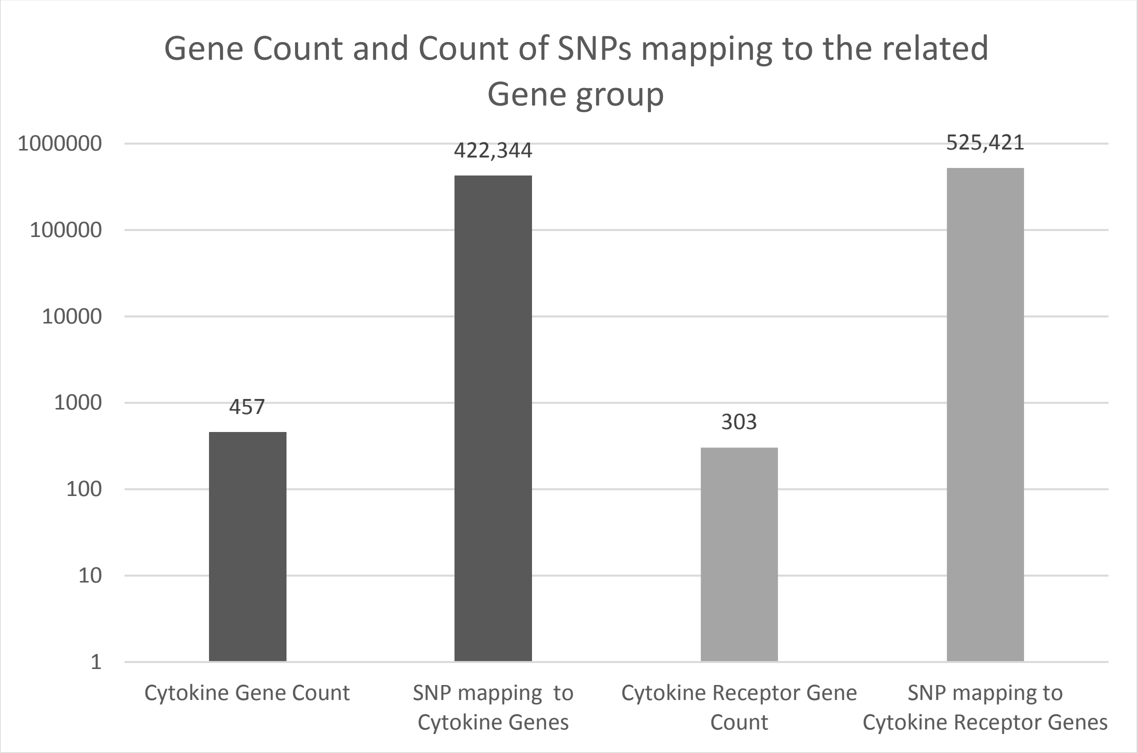
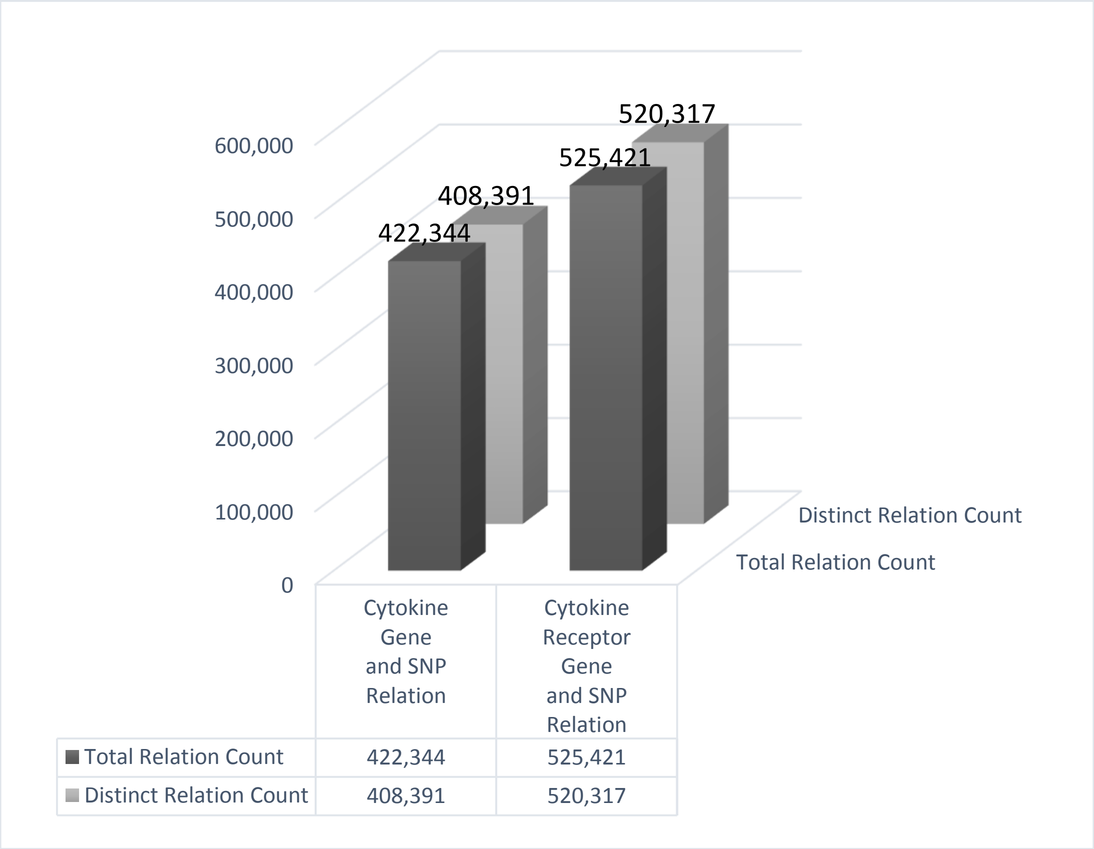


Figure 9: Distribution of SNPs mapping to genes based on the two gene groups and count of cytokine genes and cytokine receptor genes in log10. The gene count statistics presented in the figure is gathered from gene entity in CytoCAT whereas the statistics of SNPs related to genes is gathered from gene_snp entity in CytoCAT.

As mentioned before in chapter 3, there are duplicated relations between genes and SNPs because of the multiple chromosome locations. In spite of this duplications, cytokine genes and cytokine receptor genes has 96.70% and 99.03% distinct association with related SNPs, respectively. Cytokine gene and related SNPs has unique 408, 391 association. Likewise, cytokine receptor gene and related SNPs has 520, 317 unique association as seen in Figure 10.



	Cytokine Gene and SNP Relation	Cytokine Receptor Gene and SNP Relation
■ Total Relation Count	422,344	525,421
■ Distinct Relation Count	408,391	520,317

Figure 10: Total and distinct relation count of cytokine and cytokine receptor genes with SNPs according to the data stored in CytoCAT.

4.1.4. Phenotype

This entity has 121,980 records where association of 34,225 distinct rsid/markerid with phenotype is stored. 491 different experiments are done in 335,335 distinct studies to reveal these associations. As a result, 340 unique phenotype are found and annotated with distinct 450 phenotype identifier.

In this entity, 82,286 record gives information about cytokine receptor gene related SNPs and their phenotypes. Latter 39,694 entry informs about cytokine gene related SNPs and their phenotypes (Figure 11).

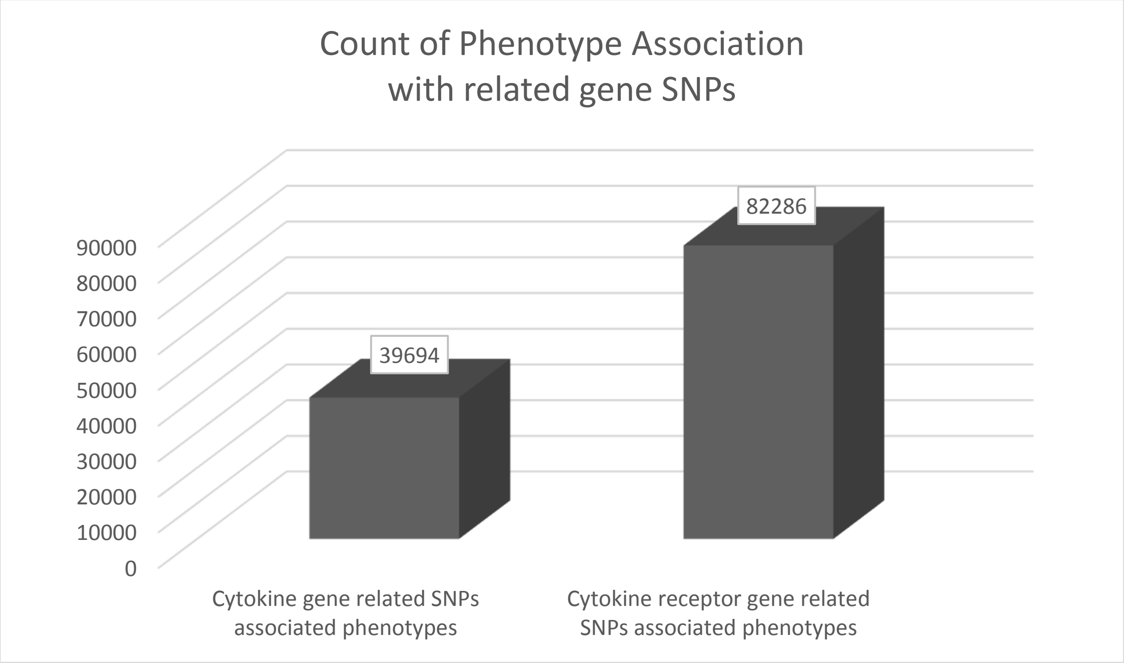


Figure 11: The count of association between phenotypes and SNPs of related group of genes recorded in CytoCAT.

4.1.5. MeSH and PhenoMesh:

Mesh entity contains 55,611 tuples. It has 27,147 unique MeSH ID assigned to 27,147 unique MeSH term. These unique MeSH ID and MeSH terms are mapped to 55, 611 unique tree IDs.

On the other hand, PhenoMesh table has 449 tuples. As most of the mesh ids are assigned to multiple phenotype ids, the table consists of 449 distinct phenotype id, however, 203 distinct mesh ID and 60 unique HPO IDs. Some of the important phenotypes in PhenoMesh table are visualized in Figure 12.

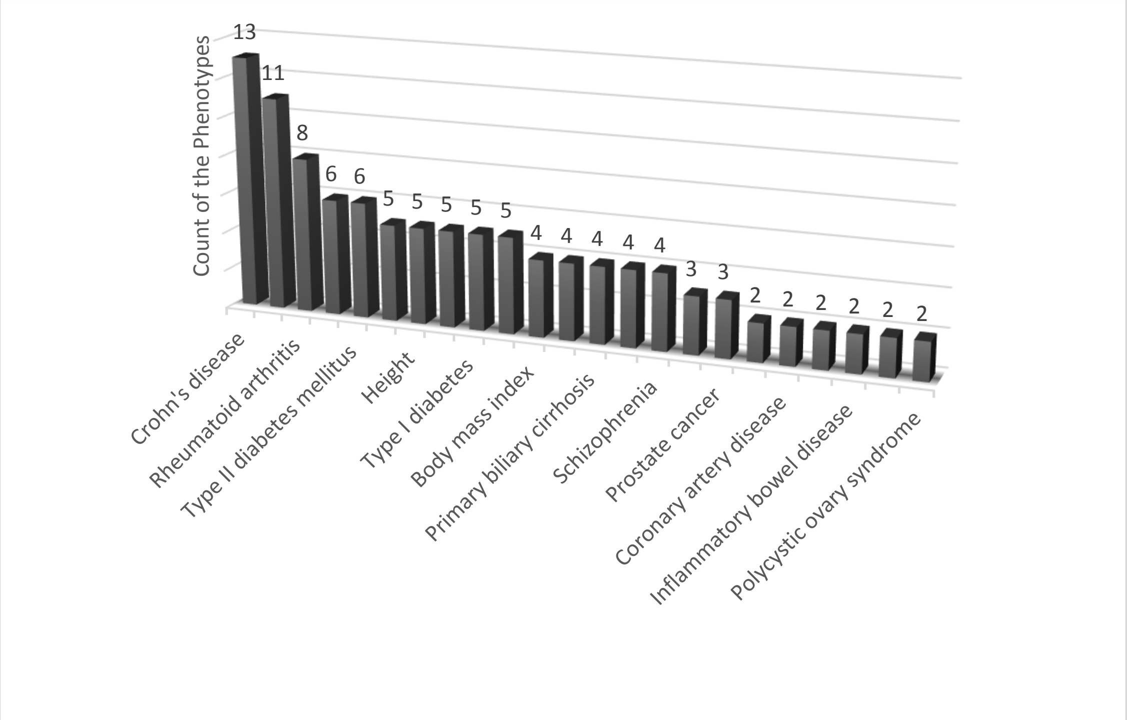


Figure 12: Some important phenotypes in PhenoMesh table and their count

We created a temporary table (pmp_temp) by joining PhenoMesh table and Mesh table with *INNER JOIN* command on meshid. This joined table allowed us to connect treeids to meshids which enabled us to categorize phenotypes of phenotype table based on the treeids. In Table 5, the summary of which MeSH categories are related to the SNPs and count of records for each eSH category in the database are given. As mentioned in chapter 2, 8 out of 16 MeSH categories are found related to the SNPs investigated in this study. Naturally, the highest count belongs to disease category.

Table 5: MeSH categories, their symbols and distinct tree ID count of each category

MeSH Category Symbol	***Count***
Anatomy[A]	47
Organism[B]	2
Diseases[C]	173
Chemicals and Drugs[D]	72
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	35
Psychiatry and Psychology[F]	15
Information Science[L]	1
Health Care[N]	6

4.1.6. Association

Association entity is the main table of CytoCAT. It contains all connections and associations between genes, SNPs located on those genes and phenotypes caused by those variations. This table is created by GENE_SNP, PHENOTYPE and the temporary table created by joining PhenoMesh and MeSH table with two inner join commands (Appendix E). There are 223092 tuples. Unique value count of each attribute is given in Table 6.

Table 6: Association entity attributes and count of the unique values for each attribute.

Attribute	**Unique Value Count**
Gene Symbol	673
Ensembl ID	673
rs ID	26850
Marker ID	26850
Phenotype ID	312
Phenotype	217
Mesh Term	146
MeSH ID	146
Tree ID	351
p-value	3716
Study ID	252
Experiment ID	351

All over the 222,747 entries, only 14.4% of them are related to cytokine genes, latter 85.6% is related to cytokine receptor genes. In Table 6, the count of distinct attributes related to the gene group is presented. According to this table, although cytokine genes are more numerous than cytokine receptor genes, number of SNP located on cytokine receptor genes is much higher than number of SNP located on cytokine genes. For that reason, the count of distinct values related to cytokine receptor genes are more than the count of distinct values related to cytokine genes for each attribute.

Table 7: The count of distinct values related to cytokine and cytokine receptor genes for each attribute in Association table

	Cytokine Related	**Cytokine Receptor Related**
Gene ID	370	303
rs Id	4095	22755
Phenotype ID	101	291
Phenotype	82	203
Mesh Term	63	140
MeSH ID	63	140
Study	89	239
Experiment	129	331
Tree ID	144	341
Total Tuple	32010	190737

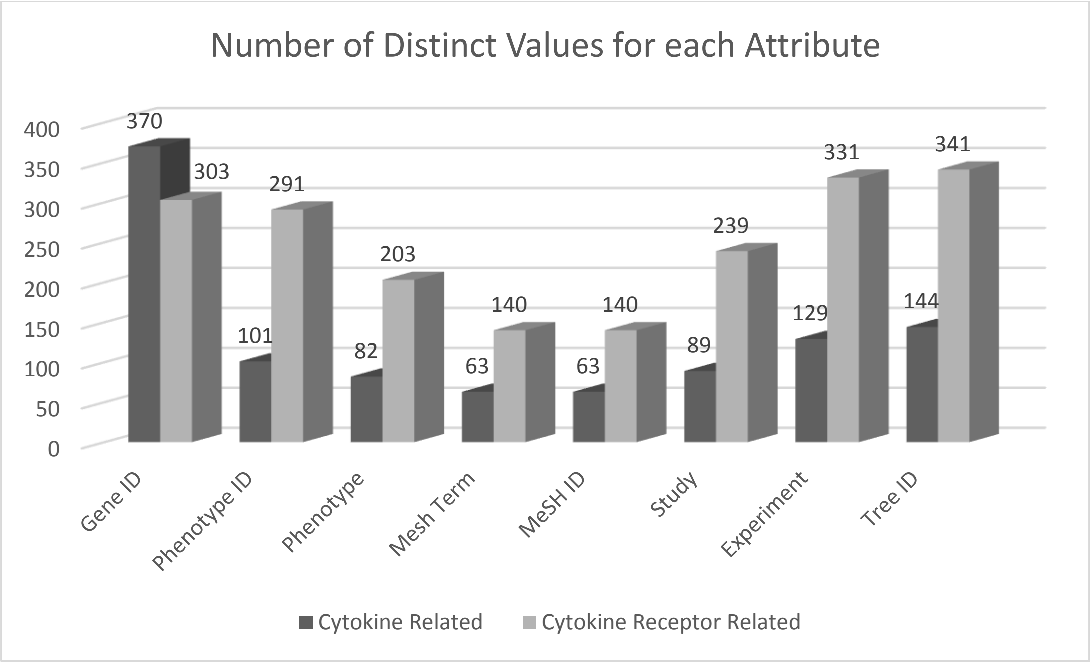


Figure 13: Number of distinct values of each attribute in Association Table

The results are analyzed according to the gene group and the MeSH tree category they are referred. Our findings are summarized with pie charts in Figure 14 and Figure 15. It can be easily seen that for both gene groups the most associated category is C: Diseases whereas second most associated category is D: Chemicals and Drugs. For other categories distribution varies between gene groups.

These association records analyzed based on their unique phenotype description. SNPs located on Cytokine genes are mapped to 82 distinct phenotype whereas SNPs located on cytokine receptor genes are associated with 204 distinct phenotype. Top 30 phenotypes which are associated with SNPs located on cytokine genes and cytokine receptor genes are presented in Figure 16 and Figure 17 respectively.

For both gene groups, ‘insulin-secreting cells’ and ‘prostatic neoplasms’ are the most associated phenotypes. It is also observed that, most of the phenotypes are major diseases. Neoplasms such as breast neoplasm, lung neoplasm, ulcerative colitis, asthma, Rheumatoid Arthritis, Crohn’s disease, cholesterol, neurological diseases like schizophrenia, Alzheimer and Parkinson disease, body mass index and insulin related disorders are all listed in the top 30 of both tables.

The associations also analyzed based on their unique MeSH term. As mentioned before in this chapter at Figure 13, there are 63 cytokine related MeSH terms and 140 cytokine receptor related MeSH terms. The list of these unique MeSH terms and their count is presented in Appendix F for cytokine related associations and in Appendix G for cytokine receptor related associations.

According to the information CytoCAT stores, the most phenotype related SNP is ‘rs1866007’. This rsid is associated 15 different MeSH term (Table 8) and located on the *RORA* (RAR-related orphan receptor A) gene. Various studies, including Zhu et al studies showed that RORA is downregulated in breast cancer (Zhu, McAvoy, Kuhn, & Smith, 2006) .

Table 8: The phenotypes rs1866007 associated and the study ID and the experiment ID these relations are observed in CytoCAT. p-values in the table are given as –log p-value.

MeSH Term	P-value	MeSH Term	P-value
Pro-insulin	2.899	Breast Neoplasms	1.282
Asthma	2.125	Parkinson Disease	1.266
Insulin-Secreting Cells	2.081	Breast Neoplasms	1.236
Insulin	1.877	Hemoglobin A, Glycosylated	1.226
Stroke	1.84	Arthritis, Rheumatoid	1.172
Insulin Resistance	1.684	Macular Degeneration	1.129
Immunoglobulin E	1.637	Blood Glucose	1.082
Brain	1.294		

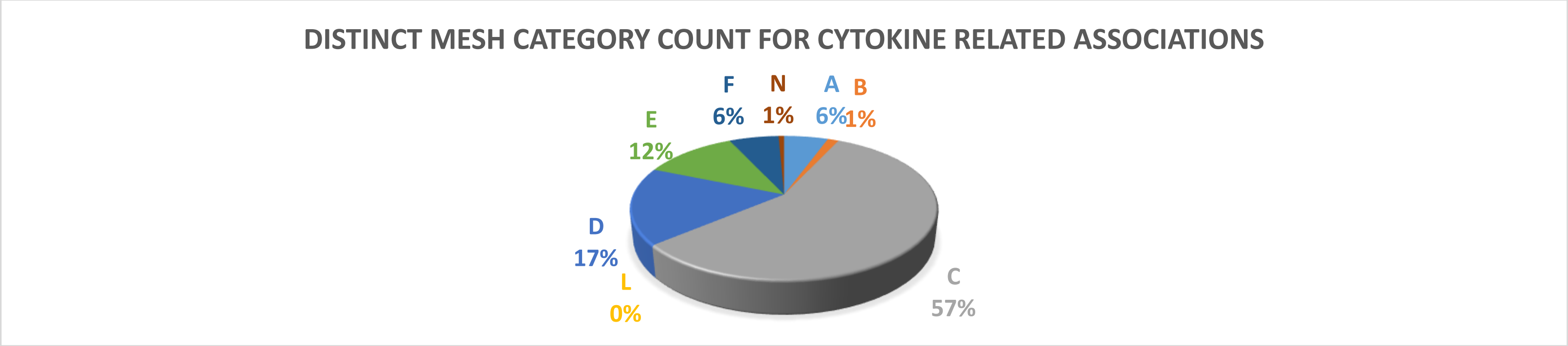


Figure 14: The distribution of cytokine related associations based on the MeSH Tree Categories. A: Anatomy, B: Organism, C: Diseases, D: Chemicals and Drugs, E: Analytical, Diagnostic and Therapeutic Techniques and Equipment, F: Psychiatry and Psychology, N: Health Care

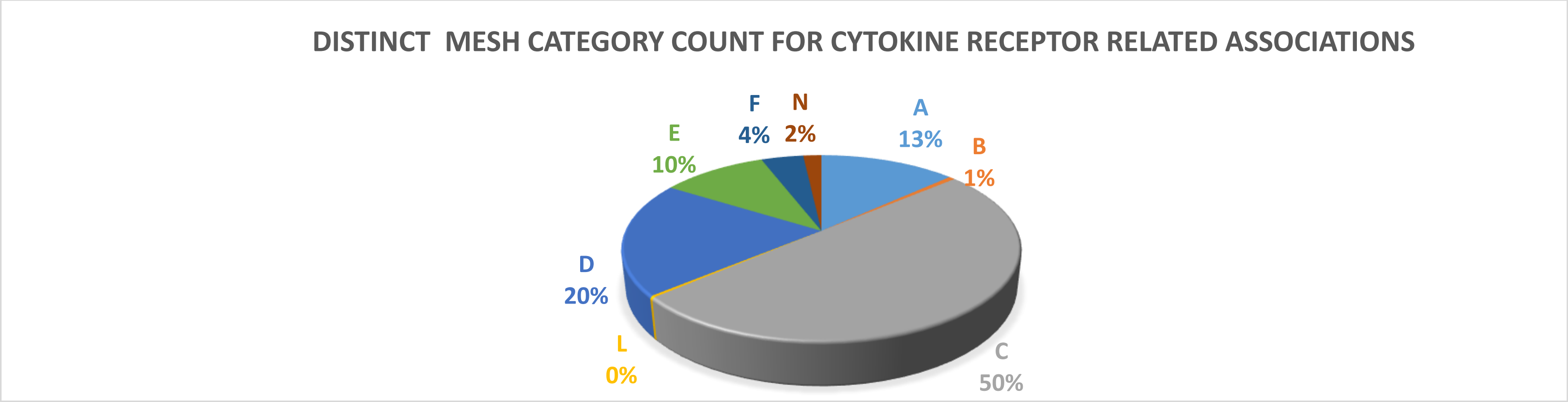


Figure 15: The distribution of cytokine receptor related associations based on the MeSH Tree Categories. A: Anatomy, B: Organism, C: Diseases, D: Chemicals and Drugs, E: Analytical, Diagnostic and Therapeutic Techniques and Equipment, F: Psychiatry and Psychology, : Information Science, N: Health Care

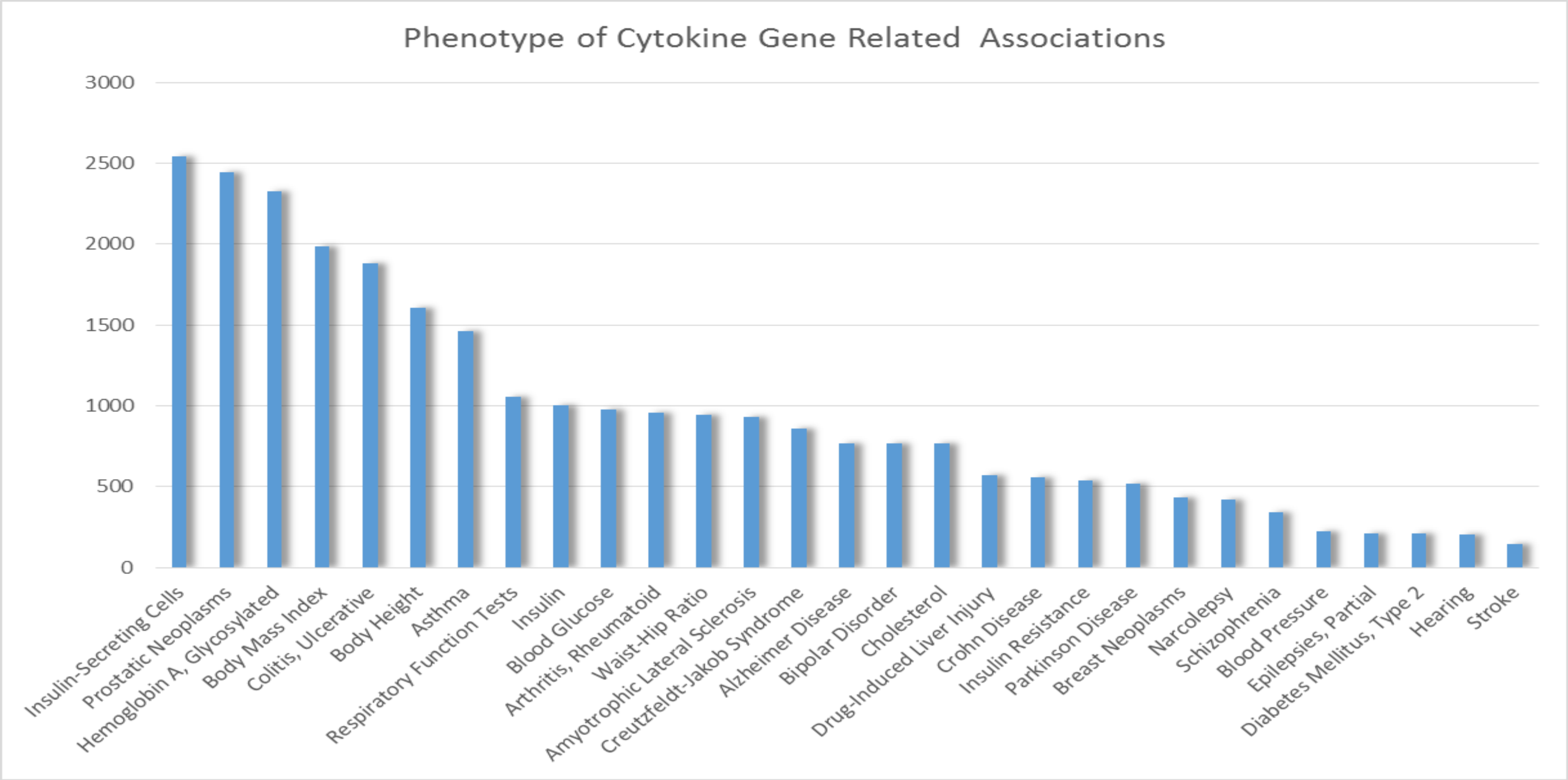


Figure 16: Top 30 phenotypes which are associated with SNPs located on Cytokine Genes.

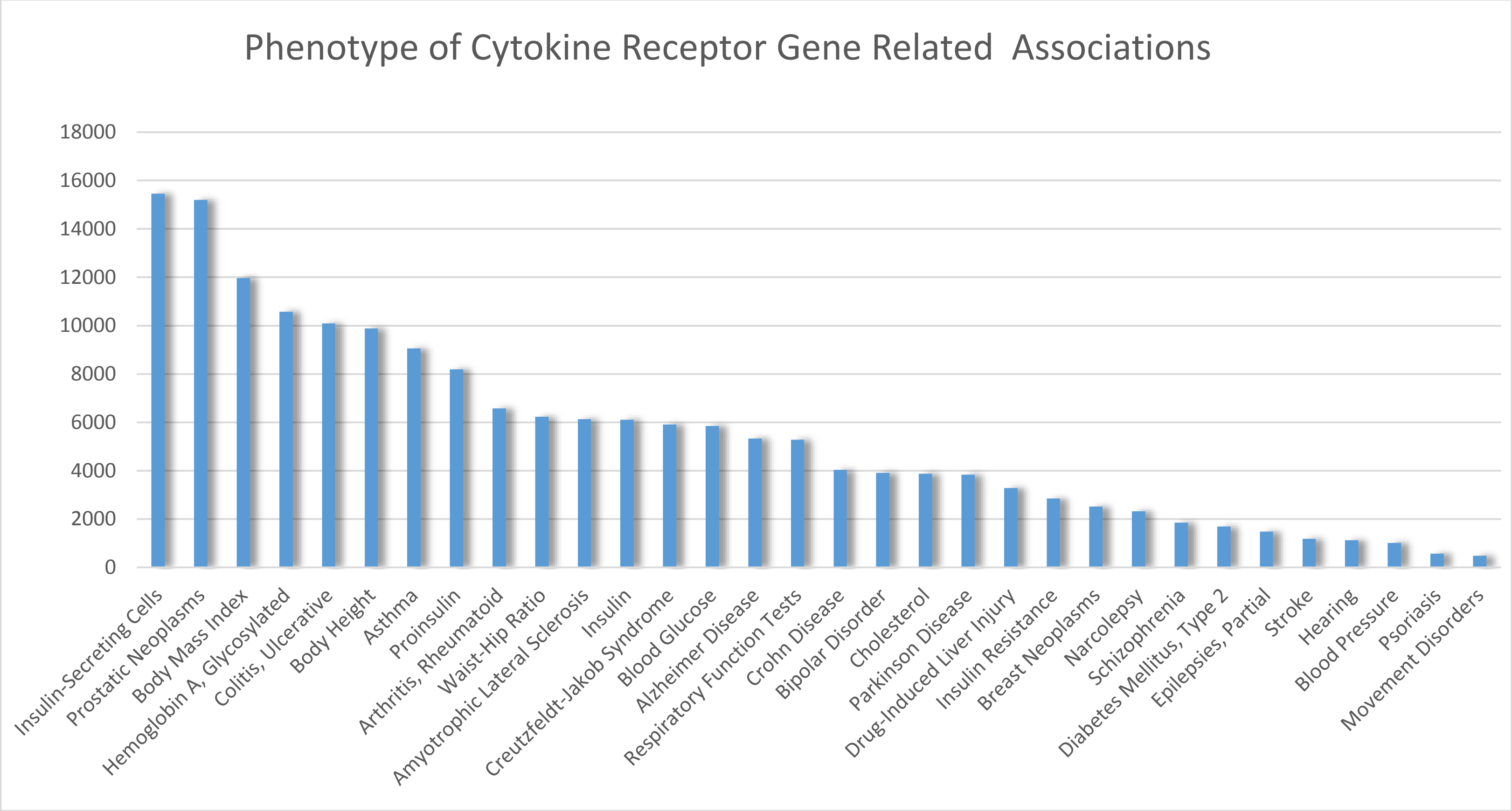


Figure 17: Top 30 phenotypes which are associated with SNPs located on Cytokine Receptor Genes

Second most phenotype associated SNPs are rs393152, rs4851004, rs4851526, rs510769, rs2416933, rs2684761, rs721862, rs7174288, rs1635291. These SNPs are all associated with 13 different phenotypes as given in Table 9 and located on cytokine receptor genes such as *CRHR1, IL18R1, IL1R2, OPRM1, NR5A1, IGF1R, NR6A1* and *RORA.*

These SNPs are associated with same phenotypes like Asthma (78%), Alzheimer disease (56%), Crohn's disease (45%), Parkinson disease (33%), Neoplasm (33%), Rheumatoid arthritis (33%). Conversely, there are unique phenotypes related to only one SNP among those 9 SNPs such as schizophrenia (rs4851526), movement disorder (rs4851004), macular degeneration (rs510769), and drug induced liver injury (rs4851526).

In CytoCAT, phenotypes are analyzed and top 30 phenotypes which have been studied the most are visualized in Figure 18. In the figure, it is observed that association of 'Body Mass Index' is studied the most which has been mentioned 4641 times. The second most studied phenotype is 'Prostate Cancer' which has been studied 4395 times. Fasting glucose related phenotypes, Ulcerative Colitis, Asthma, Rheumatoid Arthritis, Alzheimer disease, Parkinson disease, Breast Neoplasm and Schizophrenia are also listed in this top 30 phenotypes.

According to CytoCAT, Rheumatoid Arthritis is caused by 1300 different SNPs of 44 different genes. 26 of them belongs to cytokine receptor gene group such as *TGFBR2, TGFBR3, VIPR2 TNFRSF19, and TNFRS21*. Latter 18 genes belongs to cytokine gene group such as *TGFA, TGFB1, TGFB2, TNF, and SLIT1.*

In this study, number of unique SNPs associated to a phenotype is analyzed and top 20 MeSH terms which have highest number of association with different SNPs is visualized in Figure 19. In the figure, it is observed that 'Blood glucose' is the most associated term which is related to 6375 different SNPs in CytoCAT. The second 5504 different SNP. Body mass index related phenotypes, waist- hip ratios, insulin secreting cells, insulin and insulin resistance and Creutzfeldt-Jakob Syndrome are also listed in this top 20 terms.

Table 9: 9 SNPs which have highest phenotype association rank after rs1866007 and their associated phenotypes

rs393152	**rs4851004**	**rs4851526**	**rs510769**	**rs2416933**	**rs2684761**	**rs721862**	**rs7174288**	**rs1635291**
Asthma	Alzheimer Disease	Alzheimer Disease	Asthma	Amyotrophic Lateral Sclerosis	Alzheimer Disease	Alzheimer Disease	Amyotrophic Lateral Sclerosis	Alzheimer Disease
Birth Weight	Arthritis, Rheumatoid	Amyotrophic Lateral Sclerosis	Birth Weight	Bipolar Disorder	Amyotrophic Lateral Sclerosis	Amyotrophic Lateral Sclerosis	Asthma	Asthma
Blood Glucose	Asthma	Arthritis, Rheumatoid	Blood Glucose	Birth Weight	Arthritis, Rheumatoid	Bipolar Disorder	Blood Glucose	Birth Weight
Body Height	Bipolar Disorder	Asthma	Body Height	Body Height	Asthma	Birth Weight	Breast Neoplasms	Blood Glucose
Creutzfeldt-Jakob Syndrome	Body Height	Body Height	Creutzfeldt-Jakob Syndrome	Brain	Bipolar Disorder	Blood Pressure	Epilepsies, Partial	Body Height
Crohn’s Disease	Creutzfeldt-Jakob Syndrome	Colitis, Ulcerative	Epilepsies, Partial	Breast Neoplasms	Blood Glucose	Body Height	Forced Expiratory Volume	Crohn’s Disease
Epilepsies, Partial	Crohn’s Disease	Creutzfeldt-Jakob Syndrome	Hearing	Creutzfeldt-Jakob Syndrome	Body Height	Diabetes Mellitus, Type 2	Insulin Resistance	Epilepsies, Partial
Fibrinogen	Diabetic Retinopathy	Crohn’s Disease	Insulin	Diabetes Mellitus, Type 2	Colitis, Ulcerative	Glomerulonephritis, IGA	Insulin-Secreting Cells	Fibrinogen
Hearing	Drug-Induced Liver Injury	Diabetes Mellitus, Type 2	Insulin Resistance	Fibrinogen	Insulin	Hemoglobin A, Glycosylated	Koru	Hearing
Parkinson Disease	Epilepsies, Partial	Drug-Induced Liver Injury	Macular Degeneration	Glomerulonephritis, IGA	Insulin Resistance	Immunoglobulin E	Parkinson Disease	Proinsulin
Proinsulin	Hearing	Epilepsies, Partial	Parkinson Disease	Hearing	Kuru	Kuru	Proinsulin	Respiratory Function Tests
Respiratory Function Tests	Movement Disorders	Psoriasis	Stroke	Immunoglobulin E	Parkinson Disease	Narcolepsy	Prostatic Neoplasms	Stroke
Waist-Hip Ratio	Narcolepsy	Schizophrenia	Waist-Hip Ratio	Kuru	Psoriasis	Waist-Hip Ratio	Stroke	Waist-Hip Ratio

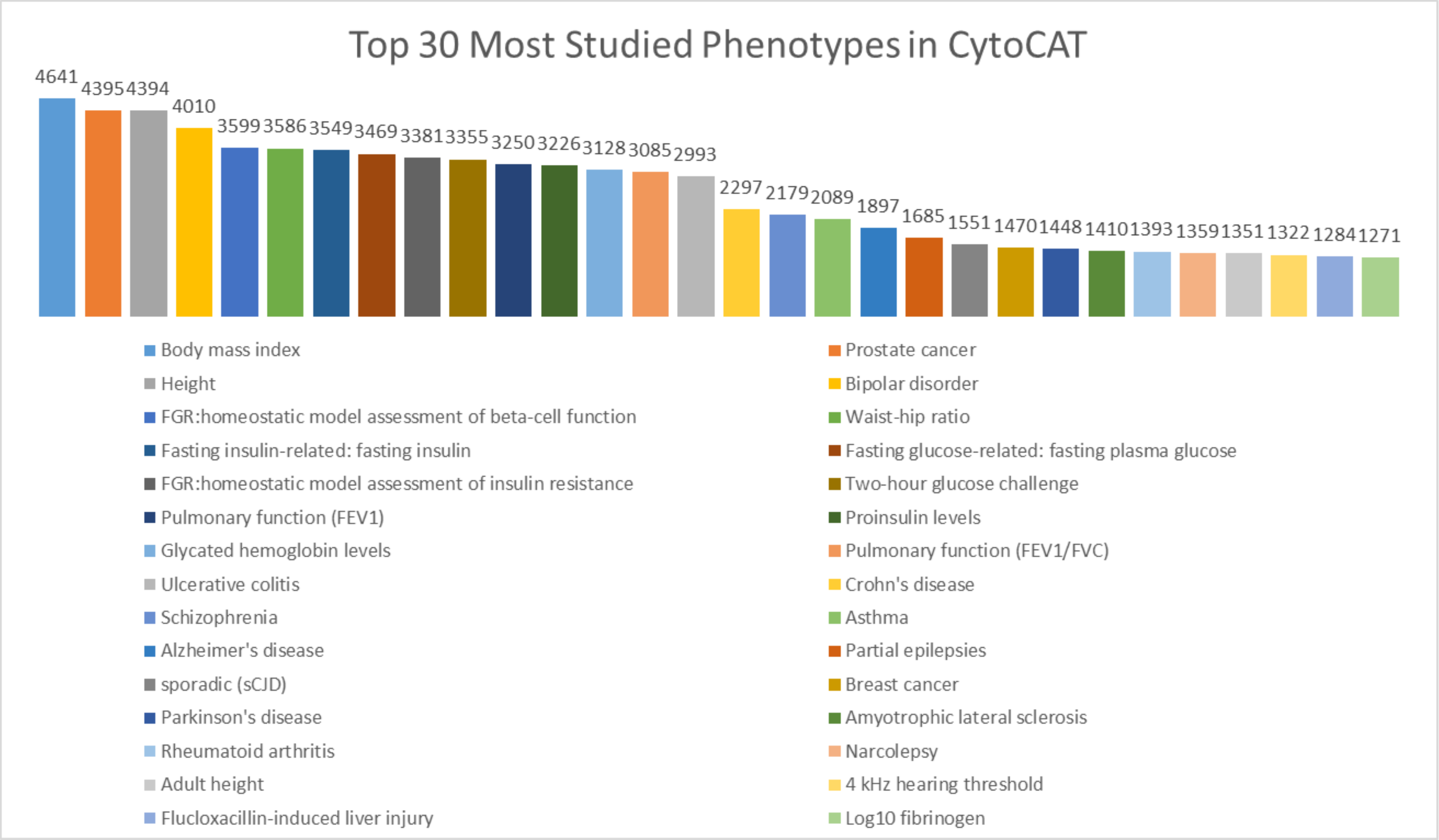


Figure 18: The top 30 phenotypes which have been studied most in CytoCAT. FGR implies fasting glucose related

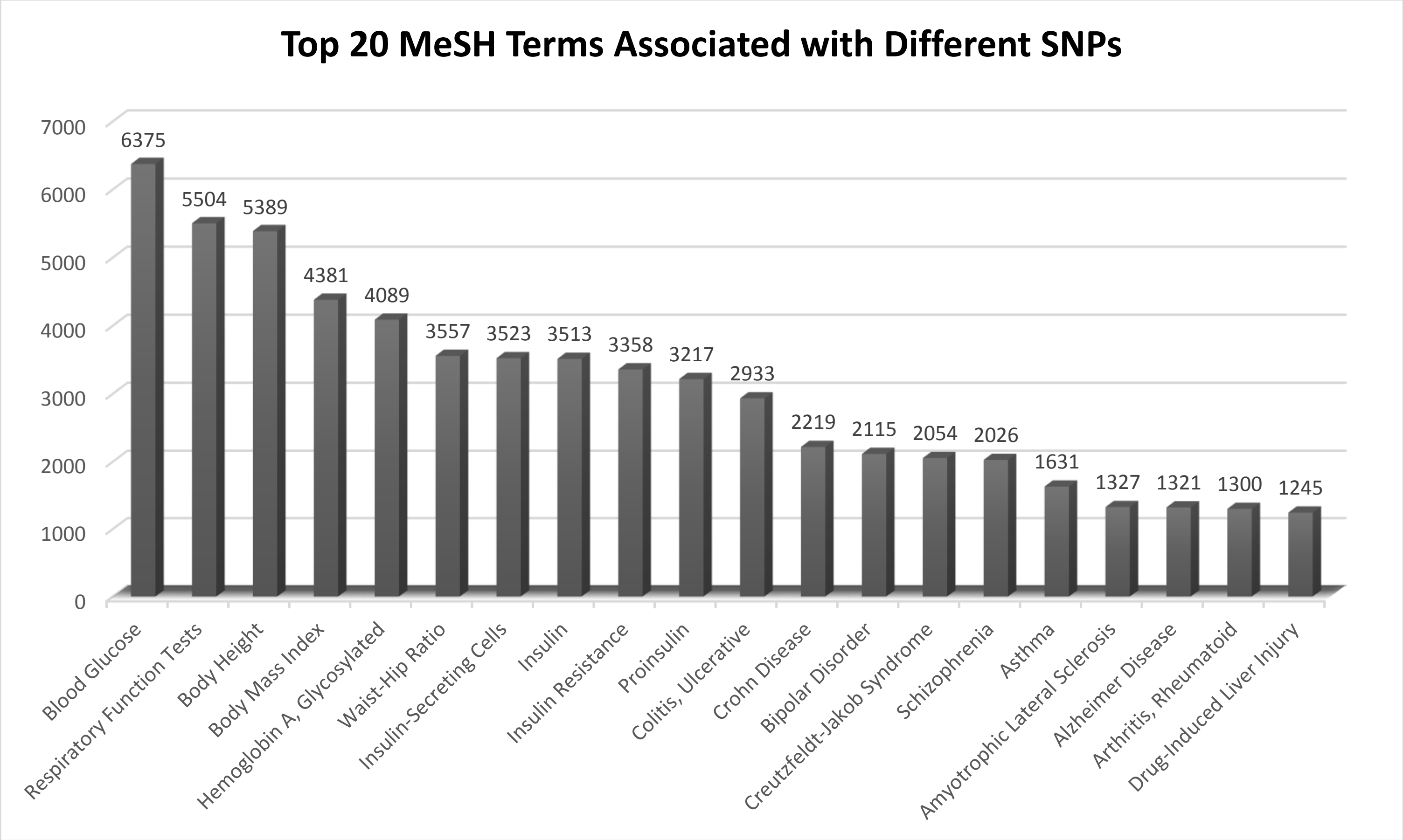


Figure 19: The top 20 MeSH terms which have highest number of association with various SNPs in CytoCAT

4.2. Case Study

4.2.1. Database of Genotypes and Phenotypes (dbGaP)

dbGaP is a public repository which is created by The National Center for Biotechnology Information (NCBI). The database stores and presents large scale genetic and phenotypic datasets of genome wide association study results. All content is arranged under four titles as study, phenotype, genotype and experiment. Every data has its own unique identifier, and prefix of these identifiers changes according to the title data belongs. For example, a study accession number starts with "phis" whereas genotype identifiers starts with "phg". These unique identifiers enables them to be cited or discussed in other publications. In public access, dbGaP provides overview of phenotype and statistics of genotypes in addition to study documents. On the other hand, authorized users get access to individual level data. (Mailman et al., 2007)

4.2.2. Cases

Jemal et. al. published the article 'Global Cancer Statics' in 2011. In the article (55), it is mentioned that based on the GLOBOCAN 2008, approximately 12.7 million cancer cases and more than 7.5 million cancer deaths are estimated to have occurred. Approximately 13% of all death in other words, 7.6 million deaths shows that cancer is a leading cause of death. According to the fact sheet published by WHO
(http://www.who.int/mediacentre/factsheets/fs297/en/index.html) the main types of cancers include lung, stomach, breast, prostate and melanoma. Therefore in this case study, prostate, lung and breast datasets are used.

When the CytoCAT analyzed, it is understood that for SNPs mapped on cytokine genes and their associated phenotypes only two types of neoplasm is observed, prostate and breast. On the other hand, SNPs mapped on cytokine receptor genes are associated with brain neoplasms, pancreatic neoplasms, lung neoplasms, nasopharyngeal neoplasms and non-small cell lung carcinoma in addition to breast neoplasms and prostatic neoplasms.

Rsid lists of three datasets from dbGaP were extracted and the ones which have unadjusted p-value higher than 0.05 are filtered out. According to this, we got 26,398 prostate neoplasm related, 27,075 breast neoplasm related, and 76,187 lung neoplasm related SNP ID from dbGaP. The count of breast neoplasm, prostatic neoplasm and lung neoplasm related rsid stored in CytoCAT is given in Table 10.

Table 10: The count of breast neoplasm, prostatic neoplasm and lung neoplasm related rsid stored in CytoCAT

Type of the Neoplasm	Cytokine Related rsid Count	Cytokine Receptor Related rsid Count	Total distinct rsid
Breast Neoplasms	117	663	780
Prostatic Neoplasms	135	812	947
Lung Neoplasms	1	3	4
Total distinct rsid	242	1390	1632

In total, 90 unique cytokine gene related SNP matched with dbGaP SNPs which is presented in Figure 20. 66 SNPs associated with breast neoplasm according to CytoCAT mapped to 44 breast neoplasms related, 17 lung neoplasms related, 5 prostate neoplasms related dbGaP SNPs. 33 SNPs associated with prostate neoplasm according to CytoCAT mapped to 4 breast neoplasms related, 25 lung neoplasms related, 4 prostate neoplasms related dbGaP SNPs.

On the other hand, 563 unique cytokine receptor gene related SNP matched with dbGaP SNPs which is presented in Figure 21. 370 SNPs associated with breast neoplasm according to CytoCAT mapped to 252 breast neoplasms related, 92 lung neoplasms related, 26 prostate neoplasms related dbGaP SNPs. 191 SNPs associated with prostate neoplasm according to CytoCAT mapped to 41 breast neoplasms related, 117 lung neoplasms related, 33 prostate neoplasms related dbGaP SNPs. 1 SNP associated with lung neoplasm and 1 SNP associated with pancreatic neoplasm associated with 1 lung neoplasm related dbGaP SNP.

4.3. CytoCAT Query Results

In CytoCAT searches can be done with three different options. By gene symbol or ensembl identifier, by dbSNP rsid or by phenotype keywords as seen in Figure 22. Besides all these optional searches, every search returns with a tabular view of gene-SNP- phenotype association, GWAS Central identifier of study and experiment the association revealed and –log p-value of association which shows significance. In addition to this gene results are linked to related Ensembl Gene page whereas dbSNP rsids are linked to related dbSNP page.

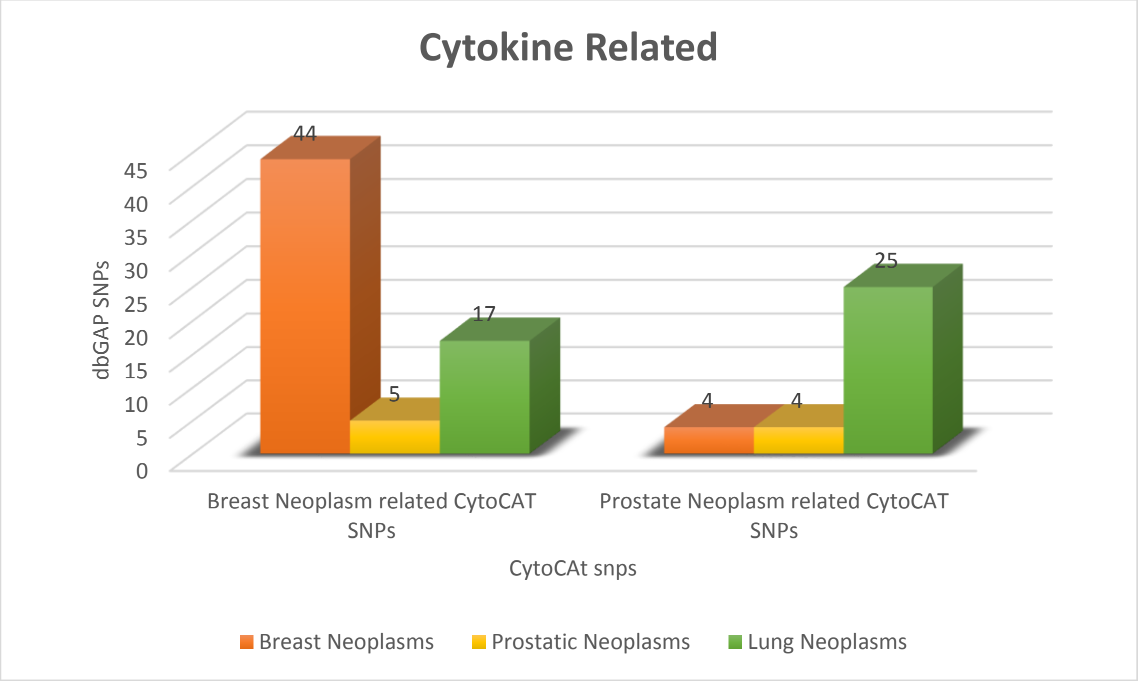


Figure 20: Cytokine gene related SNPs comparison with dbGaP SNPs for three types of neoplasm

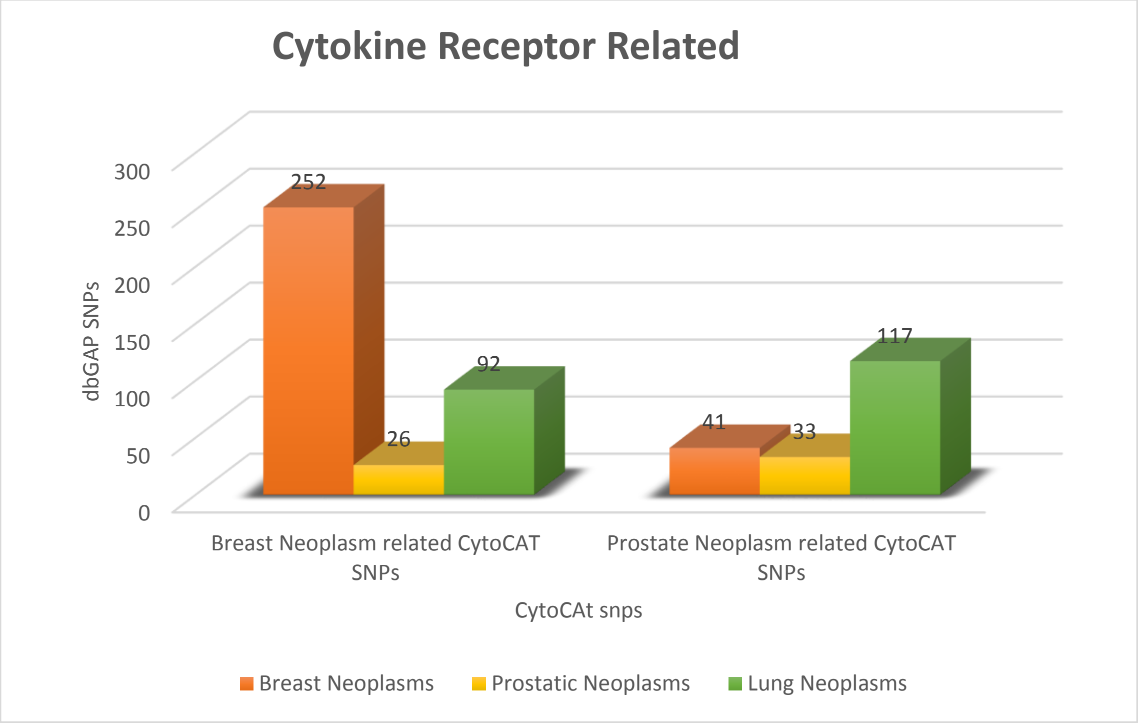


Figure 21: Cytokine receptor gene related SNPs comparison with dbGaP SNPs for three types of neoplasm

4.3.1. Search by Gene

If searches are done by option 1, via gene identifier or symbol, in result page users get total query result count, related genes Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association as shown in Figure 23 and Figure 24. In addition to this, as genes are linked to related Ensembl Gene page, users can also access general information about cytokine and cytokine receptor genes like their gene start and gene end locations, chromosome number, strand, band genes located on.

4.3.2. Search by SNP

If search is done by option 2, by dbSNP rsid, users get total query result count, related genes type, Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association is presented as presented in Figure 25. In addition to this, as dbSNP rsids are linked to related dbSNP rsid page, users can also access general information about SNPs located on cytokine and cytokine receptor genes like their varying allele, minor allele and evidence status.

4.3.3. Search by Disease

If users search by a keyword, users get total query result count, related genes type, Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association is presented as shown in Figure 26. In CytoCAT, there are two query options for "Search by Disease" as presented in Figure 27. First query box is used for direct keyword search whereas second option lists all the MeSH Disease Class tree. Users can click on the check boxes and choose multiple options from the disease class tree to see results for more than one disease class and their sub-classes.

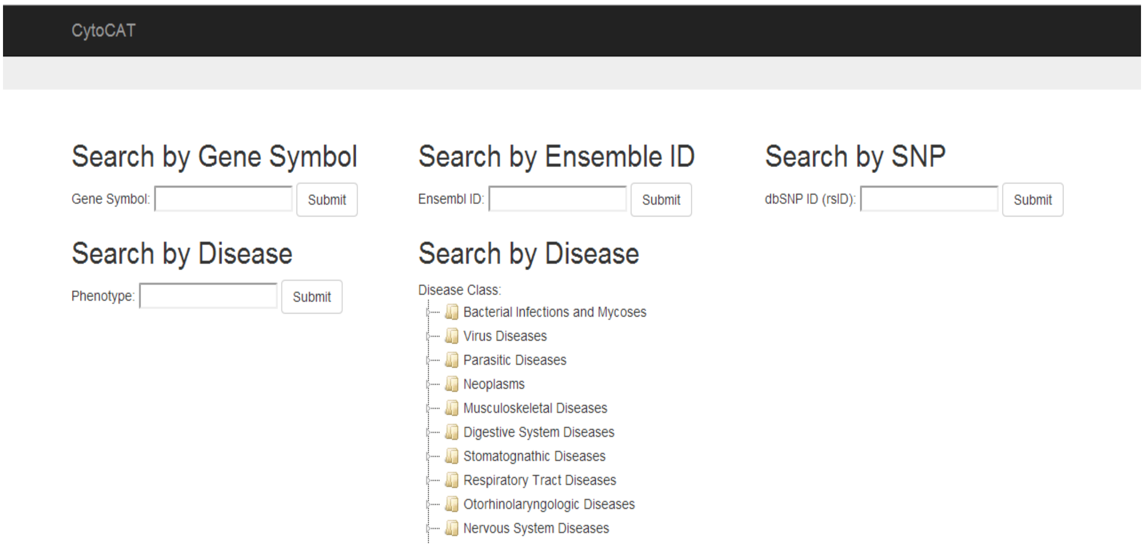
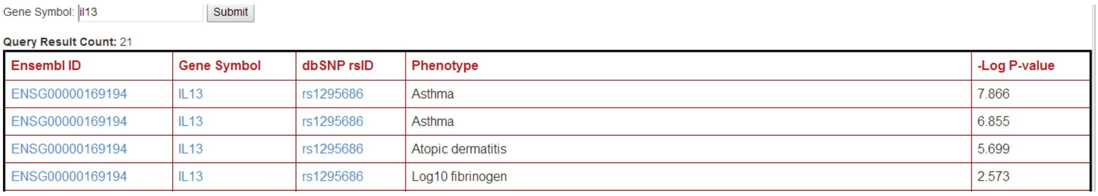


Figure 22: The main page of CytoCAT. Each CytoCAT query option can be reached through this page by users.

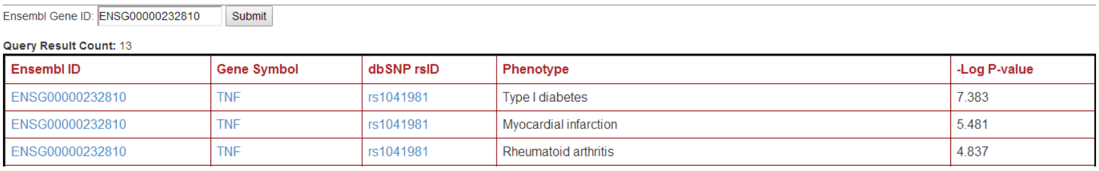


Gene Symbol: il13

Query Result Count: 21

Ensembl ID	Gene Symbol	dbSNP rsID	Phenotype	-Log P-value
ENSG00000169194	IL13	rs1295686	Asthma	7.866
ENSG00000169194	IL13	rs1295686	Asthma	6.855
ENSG00000169194	IL13	rs1295686	Atopic dermatitis	5.699
ENSG00000169194	IL13	rs1295686	Log10 fibrinogen	2.573

Figure 23: The Gene Symbol Query Result page. In this result page, total query result count, related genes Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association is presented. In addition to this, genes are linked to related Ensembl Gene page and dbSNP rsids are linked to related dbSNP rsid page.

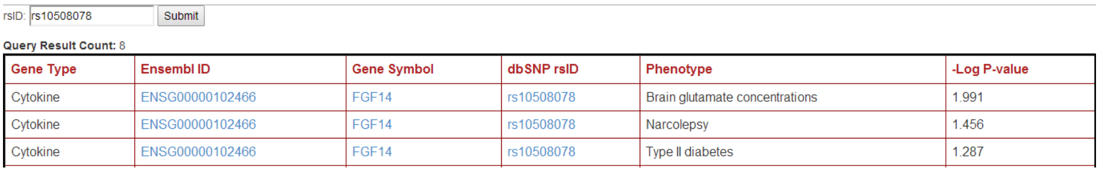


Ensembl Gene ID: ENSG00000232810

Query Result Count: 13

Ensembl ID	Gene Symbol	dbSNP rsID	Phenotype	-Log P-value
ENSG00000232810	TNF	rs1041981	Type I diabetes	7.383
ENSG00000232810	TNF	rs1041981	Myocardial infarction	5.481
ENSG00000232810	TNF	rs1041981	Rheumatoid arthritis	4.837

Figure 24: The Ensembl Gene ID Query Result page. In this result page, total query result count, related genes Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association is presented. In addition to this, genes are linked to related Ensembl Gene page and dbSNP rsids are linked to related dbSNP rsid page.

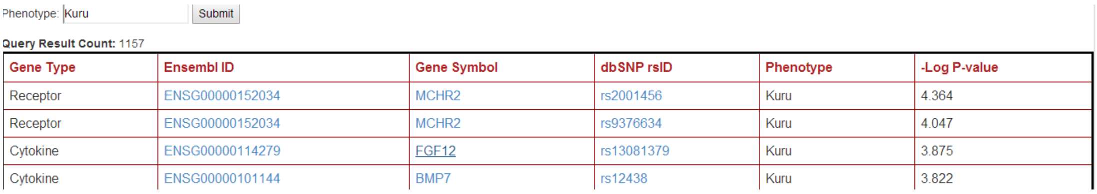


rsID: rs10508078 Submit

Query Result Count: 8

Gene Type	Ensembl ID	Gene Symbol	dbSNP rsID	Phenotype	-Log P-value
Cytokine	ENSG00000102466	FGF14	rs10508078	Brain glutamate concentrations	1.991
Cytokine	ENSG00000102466	FGF14	rs10508078	Narcolepsy	1.456
Cytokine	ENSG00000102466	FGF14	rs10508078	Type II diabetes	1.287

Figure 25: The dbSNP rsID Query Result page. In this result page, total query result count, related genes type, Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association is presented. In addition to this, genes are linked to related Ensembl Gene page and dbSNP rsids are linked to related dbSNP rsid page.

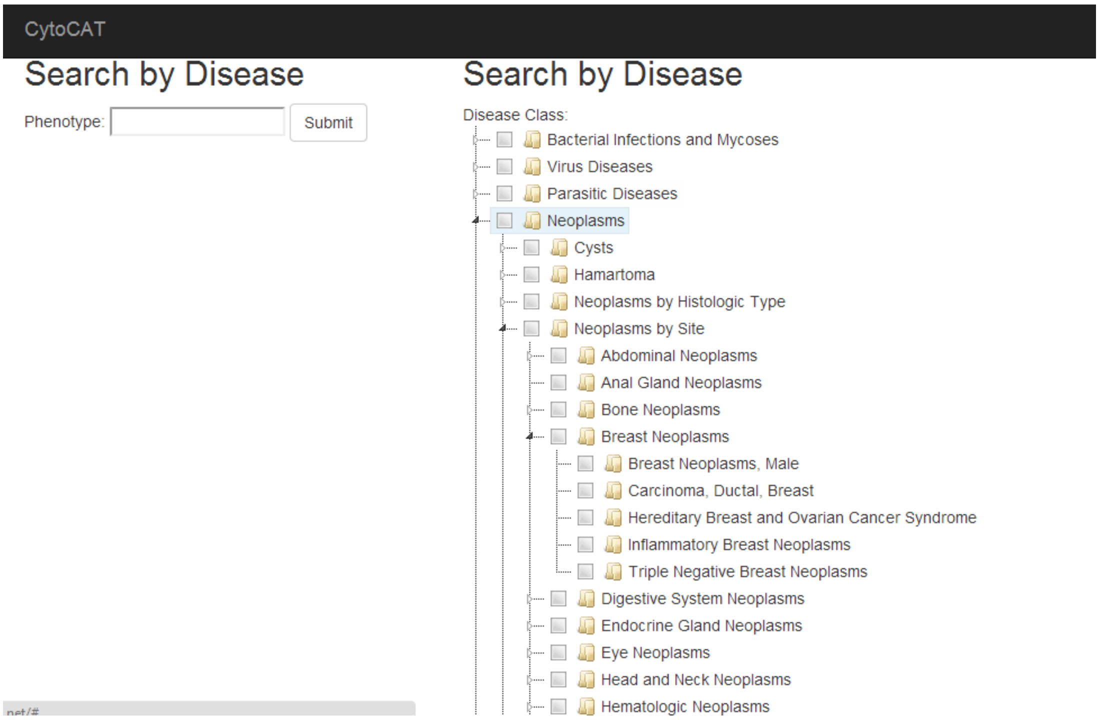


Phenotype: Kuru Submit

Query Result Count: 1157

Gene Type	Ensembl ID	Gene Symbol	dbSNP rsID	Phenotype	-Log P-value
Receptor	ENSG00000152034	MCHR2	rs2001456	Kuru	4.364
Receptor	ENSG00000152034	MCHR2	rs9376634	Kuru	4.047
Cytokine	ENSG00000114279	FGF12	rs13081379	Kuru	3.875
Cytokine	ENSG00000101144	BMP7	rs12438	Kuru	3.822

Figure 26: The Phenotype Query Result page. In this result page, total query result count, related genes type, Ensembl Gene ID, Gene Symbol, dbSNP rsid mapping on the related gene, associated phenotype and –log p-value of this association is presented. In addition to this, genes are linked to related Ensembl Gene page and dbSNP rsids are linked to related dbSNP rsid page.



Search by Disease

Phenotype: Submit

Search by Disease

Disease Class:

- □ Bacterial Infections and Mycoses
- □ Virus Diseases
- □ Parasitic Diseases
- □ Neoplasms
  - □ Cysts
  - □ Hamartoma
  - □ Neoplasms by Histologic Type
  - □ Neoplasms by Site
    - □ Abdominal Neoplasms
    - □ Anal Gland Neoplasms
    - □ Bone Neoplasms
    - □ Breast Neoplasms
      - □ Breast Neoplasms, Male
      - □ Carcinoma, Ductal, Breast
      - □ Hereditary Breast and Ovarian Cancer Syndrome
      - □ Inflammatory Breast Neoplasms
      - □ Triple Negative Breast Neoplasms
    - □ Digestive System Neoplasms
    - □ Endocrine Gland Neoplasms
    - □ Eye Neoplasms
    - □ Head and Neck Neoplasms
    - □ Hematologic Neoplasms

Figure 27: The two query options for "Search by Disease". First query box is used for direct keyword search whereas second option lists all the MeSH Disease Class tree. Users can click on the check boxes and choose multiple options from the disease class tree to see results for more than one disease class and their sub-classes.

CHAPTER 5

Discussion

5.1.1. Importance of Cytokine and Cytokine Receptors

Cytokines play an important role in the immune response by mediating the innate and adaptive immune systems. Despite their crucial role in immune response, without a receptor to bind, cytokines cannot signal, for that reason, cannot achieve their goal and function. According to this information, both cytokines and cytokine receptors play an important role in diseases can be concluded. For that reason, since last decade, scientists have been focused on cytokines and the role they play in diseases. CytoCAT, which presents the association between genes and SNPs, can be a guideline to understand the relation between genetics and diseases. Cytokinologists, immunologists, doctors and scientists who studies drug discovery and diagnosis techniques can benefit from CytoCAT, which fills the knowledge gap of disease causing genetic variations. The knowledge of disease causing genetic variations can enlighten the causes of diseases. In addition to this, it can lead scientists to identify the most appropriate drug targets which would enable more effective drug discoveries.

5.1.2. Importance of Single Nucleotide Polymorphisms

In association studies, it is seen that SNPs can function as biomarker of diseases or act as a potential disease causing genetic variation. When they function as biomarkers, disease causing genes are targeted more easily. If a SNP affects the transcription of genes to mRNA thereby affecting the coded protein, they might cause some disease related phenotypic changes. Especially non-synonymous SNPs might cause different proteins or untranslated proteins by changing or stopping the amino acid expression. For that reason, disease causing SNPs would be more informative while identifying the drug targets and as targets will be more specific, the healing power of drugs will increase. In addition to this, as SNPs are changing from population to population this catalog would also help personalized medicine studies.to summarize the importance, the knowledge of disease and SNP association will support characterization of drug targets, improvements in drug design and enhancement of drug activities.

5.1.3. Importance of CytoCAT

CytoCAT is the first catalog that presents highly annotated cytokine and cytokine receptor SNPs and their phenotype associations. Until this study, two other studies, which were mentioned in Chapter 2, have also tried to create catalogs which store the disease association of cytokine related SNPs. However, in our study, we broaden our target genes and have collated data on both cytokine and cytokine receptor genes. Rather than just focusing on diseases we also include phenotypes. Although phenotypes do not directly indicate that the patient has the disease, it might give clues about preliminary findings on diseases, which leads researchers to susceptibility prediction and early detection of diseases. For that reason, including phenotype associations in addition to disease associations increased the early detection and early diagnosis of disease, which enables early prevention of diseases.

5.1.4. Significance of the data in CytoCAT

When we analyze the stored information in CytoCAT, it is easily seen that cytokine genes are more than cytokine receptors. Despite the gene counts, both association with SNPs and phenotypes are more for cytokine receptor genes. This shows that although cytokines are the main biomolecules in immune system, they need a receptor to bind and function, thus cytokine receptors play more vital role in immunological response.

The gene and phenotype associations shows that over 765 genes, only 673 of them are related to a phenotype. 81% of cytokine genes are related to phenotypes whereas 99% of cytokine receptor genes are associated with phenotypes.

When phenotypes are mapped to MeSH tree categories, it is seen that 50% of the SNP related phenotypes are grouped under the C catalog, which defines diseases. 20% of the phenotypes are categorized under catalog D, which is comprised of drugs and chemicals. The remaining 30% is distributed among the other 6 categories. With this disease association knowledge, diseases will be detected more easily. By this way, the probability of early diagnosis will increase. Besides increasing the precise and unique drug targets, the chemicals and drug association knowledge will enhance drug design and target validation. By this way, new therapeutic approaches for many major diseases such as neoplasms and brain related diseases like Alzheimer’s disease and schizophrenia will be developed.

In the case study, 3 datasets of breast, lung and prostate from dbGaP are compared with our CytoCAT findings having breast, lung and prostate neoplasm MeSH terms. The results indicates that for both cytokine and cytokine receptor related studies show more significance in breast neoplasm associated studies. Although we expected the opposite, prostatic neoplasm studies showed low matching score. Even though lung neoplasm associated rsid count was just 4 in total, they have higher match score than prostatic neoplasm. At the end of this study, it is also seen that matching scores of cytokine receptor related SNPs are approximately five times more numerous than the cytokine related SNPs which is normal as their total count has more than a five-fold difference.

At the end of this study, we compared some of our findings with other studies. Recent studies of Liu et al., Gan-Or et.al and Edwards et al. also confirmed that rs393152, which is found as one of the most phenotype associated SNP in our study, is associated with Parkinson’s Disease (Edwards et al., 2010; Gan-Or et al., 2012; Liu et al., 2013). Another highly phenotype associated SNP we found is rs4851004 which is associated with asthma. This finding is supported by Wu et al., Galanter et al. and Spycher et al. studies (Galanter et al., 2011; Spycher et al., 2012; H. Wu et al., 2010).

CHAPTER 6

Conclusion and Future Studies

6.1. Overview

In this study, related data is extracted from different data sources and integrated to create a catalog. A relational database is developed using the PostgreSQL database management system and a web-based application is developed to store these data at a single location. This web-based application enables us to successfully share our CytoCAT catalog among researches, immunologists and doctors. This chapter presents conclusions obtained from this study and proposes further improvements for the CytoCAT.

6.2. Conclusion

In this research, our main aim was to create a catalog which would be useful in early diagnosis and prevention of diseases and help scientists to develop new therapeutic approaches for these diseases. In line with our goal, cytokine and cytokine receptor genes are used as target genes because of their broad spectrum of activity in both innate and adaptive immune system. As human genome is rich with SNPs, they are preferred as disease related phenotype biomarkers.

The gene lists are extracted from ImmPort. The association between SNPs and related genes are gathered from Ensembl Biomart where dbSNP is also used as SNP source. In addition to association information between gene and SNP, general information for each of them is also extracted from Ensembl BioMart. The most important data set, association between SNP and phenotypes, are downloaded from GWAS Central where GWAS studies are stored.

All these data sets are integrated in a database, which we named Cytokine Polymorphism Catalog (CytoCAT). PostGreSQL database management system is used via its user interface pgAdmin. The ER diagram model is designed very meticulously to minimize further database changes as maintainability is essential in database systems. In the database creation step, particular attention is paid to label entities and their attributes with clear and understandable names. In addition to these, to be able to categorize data sets in details, entities are defined with specific attributes.

As a final step, a web interface was developed to enable access to CytoCAT. While developing the user interface, ease-o-use and being basic, instructive and user-friendly is taken into consideration. Three different search options; by gene, by SNP and by keyword in catalog is given to user who will be able to find further general information about their interested genes and SNPS besides their association with phenotypes.

With this study, we created a catalog which gives both specific and general information and statistics about phenotype-gene-SNP association. At the end of this study, SNPs located on cytokine and cytokine receptor genes are revealed. In addition to this, which gene contains the highest number of SNP and which gene contains the lowest number of SNPs is also revealed. What is more, statistical comparison between SNP located on cytokine genes and SNP located on cytokine receptor genes are also brought out new information.

In addition to information of genes SNPs located on, diseases SNPs are associated are also stored in CytoCAT. The SNPs who are associated with the highest count of diseases and SNPs specific to phenotypes are all revealed. In addition to this, diseases which has the highest count of association with SNPs are also reported. Besides these, CytoCAT showed which phenotypes and diseases are studied most and which phenotypes and diseases are needed to be studied to get more information.

Apart from these specific and general information and statistics about phenotype-gene-SNP associations, CytoCAT showed drugs and chemicals affected by SNPs. Therefore, this knowledge, as we aimed, will support detection of drug target and chemical types to be used while designing the drugs and will enable researchers to early diagnose patients and to start early treatments for the ones who might develop diseases in the future. In addition to this, with this specific phenotype and SNP association, CytoCAT will help studies of personalized medicine.

At the end of this study, it is seen that 26861 different SNPs are mapped to 673 different genes and these 26861 different SNPs are associated with 217 different phenotypes and 146 different MeSH terms. These association are found as a result of 351 experiments which are conducted under 252 different studies. over 223000 associations only 14.4% of them are related to cytokine genes and SNPs located on them whereas 85.6% of these associations are related to cytokine receptor genes and SNPs located on those genes.

When phenotypes are categories based on the MeSH tree ontology, most of them are categorized under disease category which is followed by chemicals and drugs category. This showed us that, in this research we accomplished our goal which was to be a part of diagnosis and treatment processes.

6.3. Future Work

In CytoCAT, we focused on the cytokine and cytokine receptor genes and their SNPs which are effective on phenotypes. Although other studies also focused on cytokine gene polymorphisms association with disease, our study was the first that present a catalog of highly annotated cytokine receptor SNPs and their phenotype associations.

- Last but not least, to understand mechanism and the effect of SNPs on cytokine and cytokine receptors interaction, cytokine signaling pathways which are JAK-STAT pathway, MAP kinase pathway and PI3K-AKT pathway can also be analyzed.

- Another improvement of CytoCAT to enable more accurate interpretation of the CytoCAT results would be:

  - Including ethnicity of data sets
  - In addition to this, location of SNP on gene like exon, promoter can also be presented to the user.

- What is more, mutations on cytokine and cytokine receptor genes can also be included as biomarkers to find more genetic variations causing difference in phenotypes.

- Another improvement of CytoCAT would be improving the web application:

  - Firstly, an option where users can download their query results in Microsoft Excel comma separated file format would be useful.

  - Second improvement to CytoCAT would be linking results to related data sources. For example, gene identifiers and symbols could have hyperlinks to NCBI Gene site of related gene and rsids could be linked to dbSNP website.

  - Third and most important improvement would be converting CytoCAT catalog into a web-based tool.

    - Enabling users do batch queries would be the first step. As users would be able to upload a list of interested SNPs or diseases as a list in a text file, it would minimize the spend to do queries and results for each interested SNP and/or disease would be presented in one table, which in other words mean a total summary of results for the uploaded text file.

    - In addition to this, scientists working on a different subject will be able to see the intersection of their related values and cytokine and cytokine receptor SNPs, also phenotypes.

    - Also, visualization of results in a Venn diagram would facilitate the understanding of the results.

REFERENCES

Ahlers, J. D., Belyakov, I. M., Matsui, S., & Berzofsky, J. A. (2001). Mechanisms of cytokine synergy essential for vaccine protection against viral challenge. *International immunology*, *13*, 897–908.

Baran, J., Gerner, M., Haeussler, M., Nenadic, G., & Bergman, C. M. (2011). pubmed2ensembl: a resource for mining the biological literature on genes. *PloS one*, *6*(9), e24716. doi:10.1371/journal.pone.0024716

Barrett, K. E. (1996). Cytokines: sources, receptors and signalling. *Baillière's clinical gastroenterology*, *10*(1), 1–15. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8732297

Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C., & Brookes, A. J. (2013). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European journal of human genetics : EJHG*, 1–4. doi:10.1038/ejhg.2013.274

Bennett, L., Palucka, A. K., Arce, E., Cantrell, V., Borvak, J., Banchereau, J., & Pascual, V. (2003). Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *The Journal of experimental medicine*, *197*, 711–723. doi:10.1084/jem.20021553

Bhushan, S. (2011). 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 ,, (August).

Bidwell, J., Keen, L., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M. F., … D'Alfonso, S. (1999). Cytokine gene polymorphism in human disease: on-line databases. *Genes and immunity*, *1*, 3–19. doi:10.1038/sj.gene.6363645

Borish, L. C., & Steinke, J. W. (2003). 2. Cytokines and chemokines. *The Journal of allergy and clinical immunology*, *111*, S460–S475. doi:10.1067/mai.2003.108

Bowcock, A. M., Shannon, W., Du, F., Duncan, J., Cao, K., Aftergut, K., … Menter, A. (2001). Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Human molecular genetics*, *10*, 1793–1805. doi:10.1093/hmg/10.17.1793

Bower, J. E., Ganz, P. a, Irwin, M. R., Castellon, S., Arevalo, J., & Cole, S. W. (2013). Cytokine Genetic Variations and Fatigue Among Patients With Breast Cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *31*(13). doi:10.1200/JCO.2012.46.2143

Chang, J. C., & Kan, Y. W. (1979). beta 0 thalassemia, a nonsense mutation in man. *Proceedings of the National Academy of Sciences of the United States of America*, *76*, 2886–2889. doi:10.1073/pnas.76.6.2886

Chen, P. P.-S. (1976). The entity-relationship model---toward a unified view of data. *ACM Transactions on Database Systems*. doi:10.1145/320434.320440

Chen, S., Chinnaswamy, A., Biswas, S. K., Goryachev, A. B., Yap, C.-K., Lam, K.-Y., … Mishra, S. K. (2007). Cell interaction knowledgebase: an online database for innate immune cells, cytokines and chemokines. *In silico biology*, *7*, 569–574.

Cohen, M. C., & Cohen, S. (1996). Cytokine function: a study in biologic diversity. *American journal of clinical pathology*, *105*, 589–598.

Coico, R., Sunshine, G., & Benjamini, E. (2003). *Immunology: A Short Course* (5th ed.). New Jersey: John Wiley & Sons.

Commins, S. P., Borish, L., & Steinke, J. W. (2010). Immunologic messenger molecules: cytokines, interferons, and chemokines. *The Journal of allergy and clinical immunology*, *125*, S53–S72. doi:10.1016/j.jaci.2009.07.008

Costantini, S., Capone, F., Miele, M., Guerriero, E., Napolitano, M., Colonna, G., & Castello, G. (2009). CytokineDB: a database collecting biological information. *Bioinformation*, *4*, 92–93. doi:10.6026/97320630004092

Daw, E. W., Heath, S. C., & Lu, Y. (2005). Single-nucleotide polymorphism versus microsatellite markers in a combined linkage and segregation analysis of a quantitative trait. *BMC genetics*, *6 Suppl 1*, S32. doi:10.1186/1471-2156-6-S1-S32

dbSNP. (2008, 01 31). *The dbSNP Mapping Process*. Retrieved 2012, from NCBI dbSNP: http://www.ncbi.nlm.nih.gov/books/NBK44455/

Dinarello, C. A. (2000). Proinflammatory Cytokines. *CHEST*, *118*, 503–508. doi:10.1378/chest.118.2.503

Edwards, T. L., Scott, W. K., Almonte, C., Burt, A., Powell, E. H., Beecham, G. W., … Martin, E. R. (2010). Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. *Annals of human genetics*, *74*, 97–109. doi:10.1016/S1353-8020(09)70680-7

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., … Fairley, S. (2013). Ensembl 2013. *Nucleic acids research*, *41*(Database issue), D48–55. doi:10.1093/nar/gks1236

Galanter, J. M., Torgerson, D., Gignoux, C. R., Sen, S., Roth, L. A., Via, M., … Burchard, E. G. (2011). Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 Latino populations. *The Journal of allergy and clinical immunology*, *128*, 37–43.e12. doi:10.1016/j.jaci.2011.03.050

Gan-Or, Z., Bar-Shira, A., Mirelman, A., Gurevich, T., Giladi, N., & Orr-Urtreger, A. (2012). The Age at Motor Symptoms Onset in LRRK2-Associated Parkinson's Disease is Affected by a Variation in the MAPT Locus: A Possible Interaction. *Journal of Molecular Neuroscience*. doi:10.1007/s12031-011-9641-0

Gery, I., & Waksman, B. H. (1972). Potentiation of the T-lymphocyte response to mitogens. II. The cellular source of potentiating mediator(s). *The Journal of experimental medicine*, *136,* 143–155. doi:10.1084/jem.136.1.143

Guberman, J. M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., … Cutts, R. J. (2011). BioMart Central Portal: an open database network for the biological community. *Database : the journal of biological databases and curation*, *2011*, bar041. doi:10.1093/database/bar041

GWAS Central. (2013, 9 5). *Study Database Release History-Release 11*. Retrieved 2013, from Gwas Central: http://www.gwascentral.org/info/data/study-database-release-history/

Han, G.-M., Chen, S.-L., Shen, N., Ye, S., Bao, C.-D., & Gu, Y.-Y. (2003). Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray. *Genes and immunity*, *4*, 177–186. doi:10.1038/sj.gene.6363966

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., … Clamp, M. (2002). The Ensembl genome database project. *Nucleic acids research*, *30*(1), 38–41. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99161&tool=pmcentrez&rendertype=abstract

Isaacs, A., & Lindenmann, J. (1987). Virus interference. I. The interferon. By A. Isaacs and J. Lindenmann, 1957. *Journal of interferon research*, *7*, 429–438. doi:10.1098/rspb.1957.0048

Jiang, R., Wang, H., Deng, L., Hou, J., Shi, R., Yao, M., … Sun, B. (2013). IL-22 is related to development of human colon cancer by activation of STAT3. *BMC cancer*, *13*, 59. doi:10.1186/1471-2407-13-59

Kunz, M., & Ibrahim, S. M. (2009). Cytokines and cytokine profiles in human autoimmune diseases and animal models of autoimmunity. *Mediators of inflammation*, *2009*, 979258. doi:10.1155/2009/979258

Liu, J., Xiao, Q., Wang, Y., Xu, Z.-M., Yang, Q., Wang, G., … Chen, S.-D. (2013). Analysis of genome-wide association study-linked loci in Parkinson's disease of Mainland China. *Movement disorders : official journal of the Movement Disorder Society*, *28,* 1892–5. doi:10.1002/mds.25599

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., … Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*, *39*, 1181–1186. doi:10.1038/ng1007-1181

Mantovani, A., Sica, A., Sozzani, S., Allavena, P., Vecchi, A., & Locati, M. (2004). The chemokine system in diverse forms of macrophage activation and polarization. *Trends in immunology*, *25*, 677–686. doi:10.1016/j.it.2004.09.015

Marcotte, E. M., Pellegrini, M., Yeates, T. O., & Eisenberg, D. (1999). A census of protein repeats. *Journal of molecular biology*, *293*, 151–160. doi:10.1006/jmbi.1999.3136

Mier, J. W., & Gallo, R. C. (1980). Purification and some characteristics of human T-cell growth factor from phytohemagglutinin-stimulated lymphocyte-conditioned media. *Proceedings of the National Academy of Sciences of the United States of America*, *77*, 6134–6138. doi:10.1073/pnas.77.10.6134

Milano, A., Pendergrass, S. A., Sargent, J. L., George, L. K., McCalmont, T. H., Connolly, M. K., & Whitfield, M. L. (2008). Molecular subsets in the gene expression signatures of scleroderma skin. *PloS one*, *3*, e2696. doi:10.1371/journal.pone.0002696

Momjian, B. (2001). *PostgreSQL: introduction and concepts*. *Journal of digital imaging the official journal of the Society for Computer Applications in Radiology* (Vol. 22, p. 462). doi:10.1007/s10278-007-9097-5

National Library of Medicine, N. (2014). *MeSH Tree Structure 2014*. Retrieved from NLM MeSH: https://www.nlm.nih.gov/cgi/mesh/2014/MB_cgi

Opal, S. M. (2000). Anti-Inflammatory Cytokines. *Chest*, *117*, 1162–1172. doi:10.1378/chest.117.4.1162

Pevsner, J. (2009). *Bioinformatics and Functional Genomics, 2nd Ed. Tools and Applications* (p. 451). Retrieved from http://www.amazon.com/Bioinformatics-Functional-Genomics-Edition-Jonathan/dp/B004KPVA46?SubscriptionId=1V7VTJ4HA4MFT9XBJ1R2&tag=mekentosjcom-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=B004KPVA46\npapers2://publication/uuid/7BBC0F38-C8AF-4355-A446-86BB87A2081D

Rus, V., Atamas, S. P., Shustova, V., Luzina, I. G., Selaru, F., Magder, L. S., & Via, C. S. (2002). Expression of cytokine- and chemokine-related genes in peripheral blood mononuclear cells from lupus patients by cDNA array. *Clinical immunology (Orlando, Fla.)*, *102*, 283–290. doi:10.1006/clim.2001.5182

Satoh, J., Nakanishi, M., Koike, F., Miyake, S., Yamamoto, T., Kawai, M., … Yamamura, T. (2005). Microarray analysis identifies an aberrant expression of apoptosis and DNA damage-regulatory genes in multiple sclerosis. *Neurobiology of disease*, *18*, 537–550. doi:10.1016/j.nbd.2004.10.007

Sellebjerg, F., Datta, P., Larsen, J., Rieneck, K., Alsing, I., Oturai, A., … Ryder, L. P. (2008). *Gene expression analysis of interferon-beta treatment in multiple sclerosis. Multiple sclerosis (Houndmills, Basingstoke, England)* (Vol. 14, pp. 615–621). doi:10.1177/1352458507085976

Spycher, B. D., Henderson, J., Granell, R., Evans, D. M., Smith, G. D., Timpson, N. J., & Sterne, J. A. C. (2012). Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort. *Journal of Allergy and Clinical Immunology*. doi:10.1016/j.jaci.2012.06.002

Steinman, L. (2008). Nuanced roles of cytokines in three major human brain disorders. *The Journal of clinical investigation*, *118*, 3557–3563. doi:10.1172/JCI36532

Stürzebecher, S., Wandinger, K. P., Rosenwald, A., Sathyamoorthy, M., Tzou, A., Mattar, P., … McFarland, H. F. (2003). Expression profiling identifies responder and non-responder phenotypes to interferon-beta in multiple sclerosis. *Brain : a journal of neurology*, *126*, 1419–1429. doi:10.1093/brain/awg147

Su, L. F. (2008). Updates on high-throughput molecular profiling for the study of rheumatoid arthritis. *The Israel Medical Association journal : IMAJ*, *10*, 307–309.

Tan, F. K., Hildebrand, B. A., Lester, M. S., Stivers, D. N., Pounds, S., Zhou, X., … Arnett, F. C. (2005). Classification analysis of the transcriptosome of nonlesional cultured dermal fibroblasts from systemic sclerosis patients with early disease. *Arthritis and rheumatism*, *52*, 865–876. doi:10.1002/art.20871

The ImmPort. (2009). *About ImmPort*. Retrieved from The Immunology Database and Analysis Portal (ImmPort): https://immport.niaid.nih.gov

Turnpenny, P. D., & Ellard, S. (2011). Emery's elements of medical genetics. *Churchill Livingstone*.

Van der Pouw Kraan, T. C. T. M., van Gaalen, F. A., Kasperkovitz, P. V, Verbeet, N. L., Smeets, T. J. M., Kraan, M. C., … Verweij, C. L. (2003). Rheumatoid arthritis is a heterogeneous disease: evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues. *Arthritis and rheumatism*, *48*, 2132–2145. doi:10.1002/art.11096

Verrecchia, F., & Mauviel, A. (2004). TGF-beta and TNF-alpha: antagonistic cytokines controlling type I collagen gene expression. *Cellular signalling*, *16*, 873–880. doi:10.1016/S0898-6568(04)00030-0

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics, selection, evolution : GSE*, *34*, 275–305. doi:10.1186/1297-9686-34-3-275

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human mutation*, *17*, 263–270. doi:10.1002/humu.22

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. *Nature*, *171*, 737–738. doi:10.1097/BLO.0b013e31814b9304

Whitfield, M. L., Finlay, D. R., Murray, J. I., Troyanskaya, O. G., Chi, J.-T., Pergamenschikov, A., … Connolly, M. K. (2003). Systemic and cell type-specific gene expression patterns in scleroderma skin. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 12319–12324. doi:10.1073/pnas.1635114100

Wood, P. (2006). *Understanding Immunology* (2nd ed.). Essex: Pearson Education Limited.

Wu, H., Romieu, I., Shi, M., Hancock, D. B., Li, H., Sienra-Monge, J.-J., … London, S. J. (2010). Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *The Journal of allergy and clinical immunology*, *125*, 321–327.e13. doi:10.1016/j.jaci.2009.09.007

Wu, J., & Jiang, R. (2013). Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *TheScientificWorldJournal*, *2013*, 675851. doi:10.1155/2013/675851

Zhou, X., Krueger, J. G., Kao, M.-C. J., Lee, E., Du, F., Menter, A., … Bowcock, A. M. (2003). Novel mechanisms of T-cell and dendritic cell activation revealed by profiling of psoriasis on the 63,100-element oligonucleotide array. *Physiological genomics*, *13*, 69–78. doi:10.1152/physiolgenomics.00157.2002

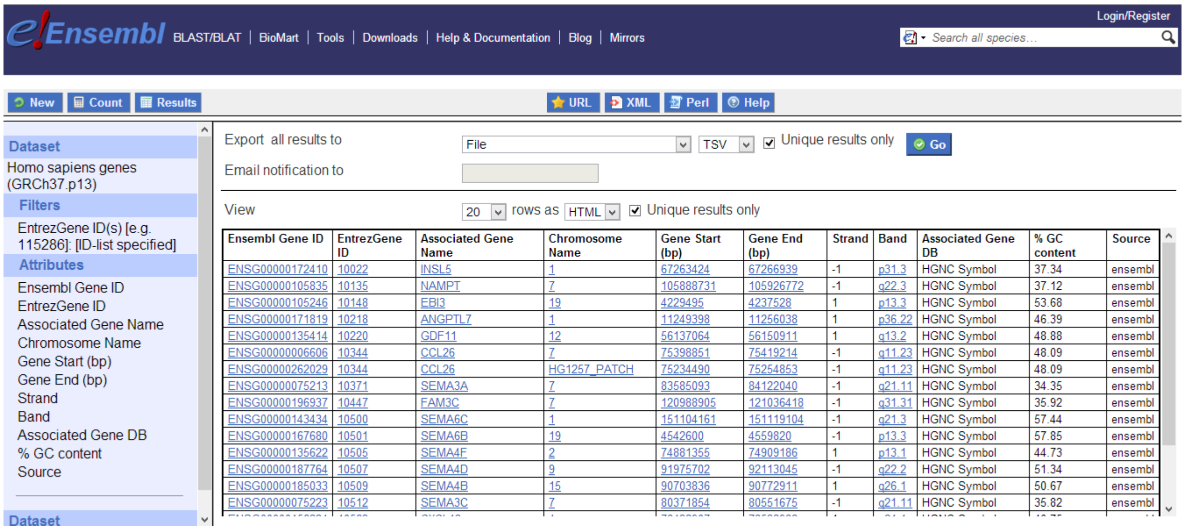
Zhu, Y., McAvoy, S., Kuhn, R., & Smith, D. I. (2006). RORA, a large common fragile site gene, is involved in cellular stress response. *Oncogene*, *25*, 2901–2908. doi:10.1038/sj.onc.1209314

Zienolddiny, S., & Skaug, V. (2012). Single nucleotide polymorphisms as susceptibility, prognostic, and therapeutic markers of nonsmall cell lung cancer. *Lung Cancer: Targets and Therapy*, *3*, 1–14. Retrieved from http://search.proquest.com/professional/docview/1032568508?accountid=138535\nhttp://www.dovepress.com/getfile.php?fileID=11728 LA - eng

APPENDICES

APPENDIX A: CONVERSION QUERIES OF ENTREZ GENE IDENTIFIERS TO ENSEMBL GENE IDENTIFIERS.

a) For cytokine gene list


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
New
Count
Results
URL
XML
Perl
Help
Go

Dataset

Homo sapiens genes (GRCh37.p13)

Filters

EntrezGene ID(s) [e.g. 115286]: [ID-list specified]

Attributes

Ensembl Gene ID
EntrezGene ID
Associated Gene Name
Chromosome Name
Gene Start (bp)
Gene End (bp)
Strand
Band
Associated Gene DB
% GC content
Source

Dataset

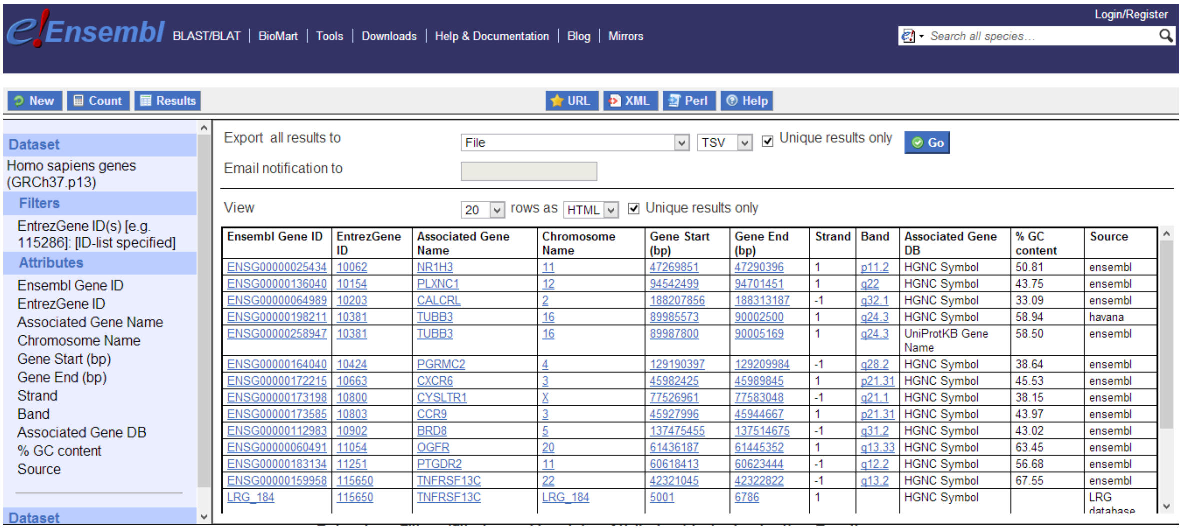
Export all results to File TSV ☑ Unique results only

Email notification to

View 20 rows as HTML ☑ Unique results only

Ensembl Gene ID	EntrezGene ID	Associated Gene Name	Chromosome Name	Gene Start (bp)	Gene End (bp)	Strand	Band	Associated Gene DB	% GC content	Source
ENSG00000172410	10022	INSL5	1	67263424	67266939	-1	p31.3	HGNC Symbol	37.34	ensembl
ENSG00000105835	10135	NAMPT	7	105888731	105926772	-1	q22.3	HGNC Symbol	37.12	ensembl
ENSG00000105246	10148	EBI3	19	4229495	4237528	1	p13.3	HGNC Symbol	53.68	ensembl
ENSG00000171819	10218	ANGPTL7	1	11249398	11256038	1	p36.22	HGNC Symbol	46.39	ensembl
ENSG00000135414	10220	GDF11	12	56137064	56150911	1	q13.2	HGNC Symbol	48.88	ensembl
ENSG00000006606	10344	CCL26	7	75398851	75419214	-1	q11.23	HGNC Symbol	48.09	ensembl
ENSG00000262029	10344	CCL26	HG1257_PATCH	75234490	75254853	-1	q11.23	HGNC Symbol	48.09	ensembl
ENSG00000075213	10371	SEMA3A	7	83585093	84122040	-1	q21.11	HGNC Symbol	34.35	ensembl
ENSG00000196937	10447	FAM3C	7	120988905	121036418	-1	q31.31	HGNC Symbol	35.92	ensembl
ENSG00000143434	10500	SEMA6C	1	151104161	151119104	-1	q21.3	HGNC Symbol	57.44	ensembl
ENSG00000167680	10501	SEMA6B	19	4542600	4559820	-1	p13.3	HGNC Symbol	57.85	ensembl
ENSG00000135622	10505	SEMA4F	2	74881355	74909186	1	p13.1	HGNC Symbol	44.73	ensembl
ENSG00000187764	10507	SEMA4D	9	91975702	92113045	-1	q22.2	HGNC Symbol	51.34	ensembl
ENSG00000185033	10509	SEMA4B	15	90703836	90772911	1	q26.1	HGNC Symbol	50.67	ensembl
ENSG00000075223	10512	SEMA3C	7	80371854	80551675	-1	q21.11	HGNC Symbol	35.82	ensembl

b) For cytokine receptor gene list.


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
New
Count
Results
URL
XML
Perl
Help
Go

Dataset

Homo sapiens genes (GRCh37.p13)

Filters

EntrezGene ID(s) [e.g. 115286]: [ID-list specified]

Attributes

Ensembl Gene ID
EntrezGene ID
Associated Gene Name
Chromosome Name
Gene Start (bp)
Gene End (bp)
Strand
Band
Associated Gene DB
% GC content
Source

Dataset

Export all results to File TSV ☑ Unique results only

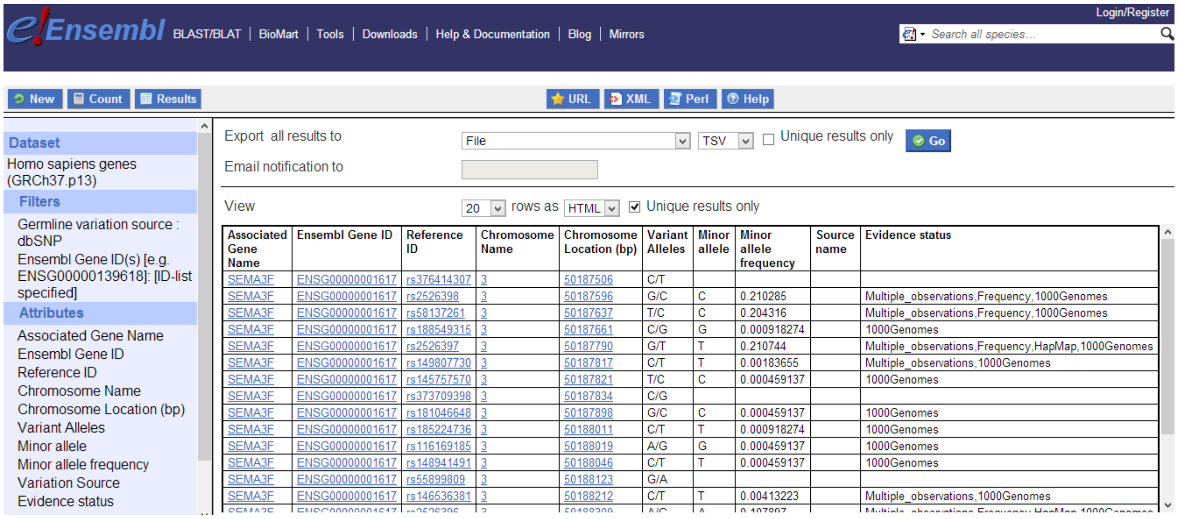
Email notification to

View 20 rows as HTML ☑ Unique results only

Ensembl Gene ID	EntrezGene ID	Associated Gene Name	Chromosome Name	Gene Start (bp)	Gene End (bp)	Strand	Band	Associated Gene DB	% GC content	Source
ENSG00000025434	10062	NR1H3	11	47269851	47290396	1	p11.2	HGNC Symbol	50.81	ensembl
ENSG00000136040	10154	PLXNC1	12	94542499	94701451	1	q22	HGNC Symbol	43.75	ensembl
ENSG00000064989	10203	CALCRL	2	188207856	188313187	-1	q32.1	HGNC Symbol	33.09	ensembl
ENSG00000198211	10381	TUBB3	16	89985573	90002500	1	q24.3	HGNC Symbol	58.94	havana
ENSG00000258947	10381	TUBB3	16	89987800	90005169	1	q24.3	UniProtKB Gene Name	58.50	ensembl
ENSG00000164040	10424	PGRMC2	4	129190397	129209984	-1	q28.2	HGNC Symbol	38.64	ensembl
ENSG00000172215	10663	CXCR6	3	45982425	45989845	1	p21.31	HGNC Symbol	45.53	ensembl
ENSG00000173198	10800	CYSLTR1	X	77526961	77583048	-1	q21.1	HGNC Symbol	38.15	ensembl
ENSG00000173585	10803	CCR9	3	45927996	45944667	1	p21.31	HGNC Symbol	43.97	ensembl
ENSG00000112983	10902	BRD8	5	137475455	137514675	-1	q31.2	HGNC Symbol	43.02	ensembl
ENSG00000060491	11054	OGFR	20	61436187	61445352	1	q13.33	HGNC Symbol	63.45	ensembl
ENSG00000183134	11251	PTGDR2	11	60618413	60623444	-1	q12.2	HGNC Symbol	56.68	ensembl
ENSG00000159958	115650	TNFRSF13C	22	42321045	42322822	-1	q13.2	HGNC Symbol	67.55	ensembl
LRG_184	115650	TNFRSF13C	LRG_184	5001	6786	1		HGNC Symbol		LRG database

APPENDIX B: ASSOCIATION QUERIES BETWEEN ENSEMBL GENE IDENTIFIERS AND DBSNP IDENTIFIERS.

a) For cytokine gene list



Dataset

Homo sapiens genes (GRCh37.p13)

Filters

Germline variation source : dbSNP

Ensembl Gene ID(s) [e.g. ENSG00000139618]: [ID-list specified]

Attributes

Associated Gene Name
Ensembl Gene ID
Reference ID
Chromosome Name
Chromosome Location (bp)
Variant Alleles
Minor allele
Minor allele frequency
Variation Source
Evidence status

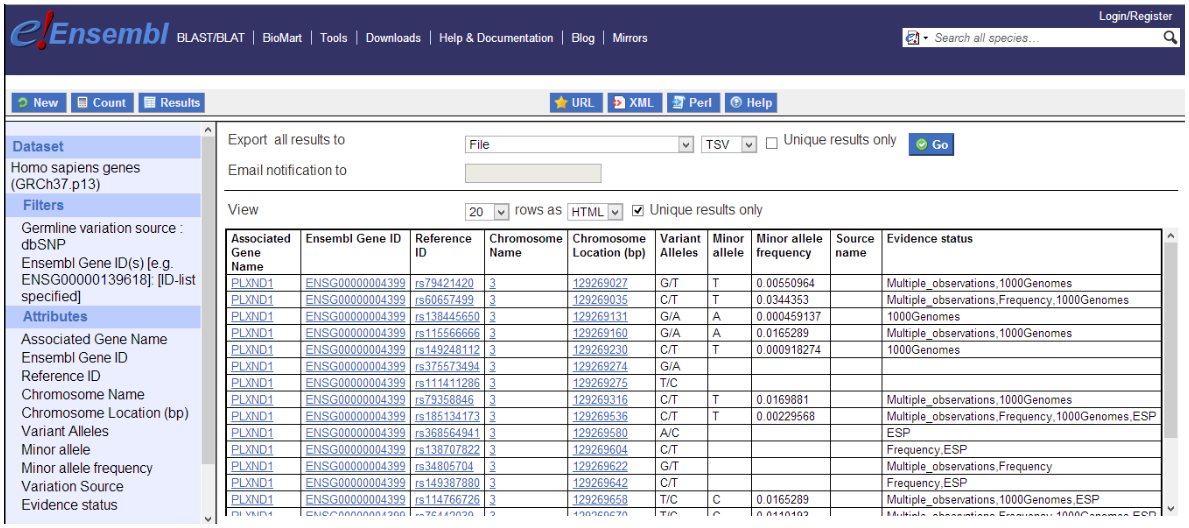
Export all results to File TSV □ Unique results only

Email notification to

View 20 rows as HTML ☑ Unique results only

Associated Gene Name	Ensembl Gene ID	Reference ID	Chromosome Name	Chromosome Location (bp)	Variant Alleles	Minor allele	Minor allele frequency	Source name	Evidence status
SEMA3F	ENSG0000001617	rs376414307	3	50187506	C/T				
SEMA3F	ENSG0000001617	rs2526398	3	50187596	G/C	C	0.210285		Multiple_observations,Frequency,1000Genomes
SEMA3F	ENSG0000001617	rs58137261	3	50187637	T/C	C	0.204316		Multiple_observations,Frequency,1000Genomes
SEMA3F	ENSG0000001617	rs188549315	3	50187661	C/G	G	0.000918274		1000Genomes
SEMA3F	ENSG0000001617	rs2526397	3	50187790	G/T	T	0.210744		Multiple_observations,Frequency,HapMap,1000Genomes
SEMA3F	ENSG0000001617	rs149807730	3	50187817	C/T	T	0.00183655		Multiple_observations,1000Genomes
SEMA3F	ENSG0000001617	rs145757570	3	50187821	T/C	C	0.000459137		1000Genomes
SEMA3F	ENSG0000001617	rs373709398	3	50187834	C/G				
SEMA3F	ENSG0000001617	rs181046648	3	50187898	G/C	C	0.000459137		1000Genomes
SEMA3F	ENSG0000001617	rs185224736	3	50188011	C/T	T	0.000918274		1000Genomes
SEMA3F	ENSG0000001617	rs116169185	3	50188019	A/G	G	0.000459137		1000Genomes
SEMA3F	ENSG0000001617	rs148941491	3	50188046	C/T	T	0.000459137		1000Genomes
SEMA3F	ENSG0000001617	rs55899809	3	50188123	G/A				
SEMA3F	ENSG0000001617	rs146536381	3	50188212	C/T	T	0.00413223		Multiple_observations,1000Genomes

b) For cytokine receptor gene list.



Dataset

Homo sapiens genes (GRCh37.p13)

Filters

Germline variation source : dbSNP

Ensembl Gene ID(s) [e.g. ENSG00000139618]: [ID-list specified]

Attributes

Associated Gene Name
Ensembl Gene ID
Reference ID
Chromosome Name
Chromosome Location (bp)
Variant Alleles
Minor allele
Minor allele frequency
Variation Source
Evidence status

Export all results to File TSV □ Unique results only

Email notification to

View 20 rows as HTML ☑ Unique results only

Associated Gene Name	Ensembl Gene ID	Reference ID	Chromosome Name	Chromosome Location (bp)	Variant Alleles	Minor allele	Minor allele frequency	Source name	Evidence status
PLXND1	ENSG00000004399	rs79421420	3	129269027	G/T	T	0.00550964		Multiple_observations,1000Genomes
PLXND1	ENSG00000004399	rs60657499	3	129269035	C/T	T	0.0344353		Multiple_observations,Frequency,1000Genomes
PLXND1	ENSG00000004399	rs138445650	3	129269131	G/A	A	0.000459137		1000Genomes
PLXND1	ENSG00000004399	rs115566666	3	129269160	G/A	A	0.0165289		Multiple_observations,1000Genomes
PLXND1	ENSG00000004399	rs149248112	3	129269230	C/T	T	0.000918274		1000Genomes
PLXND1	ENSG00000004399	rs375573494	3	129269274	G/A				
PLXND1	ENSG00000004399	rs111411286	3	129269275	T/C				
PLXND1	ENSG00000004399	rs79358846	3	129269316	C/T	T	0.0169881		Multiple_observations,1000Genomes
PLXND1	ENSG00000004399	rs185134173	3	129269536	C/T	T	0.00229568		Multiple_observations,Frequency,1000Genomes,ESP
PLXND1	ENSG00000004399	rs368564941	3	129269580	A/C				ESP
PLXND1	ENSG00000004399	rs138707822	3	129269604	C/T				Frequency,ESP
PLXND1	ENSG00000004399	rs34805704	3	129269622	G/T				Multiple_observations,Frequency
PLXND1	ENSG00000004399	rs149387880	3	129269642	C/T				Frequency,ESP
PLXND1	ENSG00000004399	rs114766726	3	129269658	T/C	C	0.0165289		Multiple_observations,1000Genomes,ESP

APPENDIX C: PYTHON SCRIPT TO EXTRACT INFORMATION FROM 2014 MESH TREES

```
#!/usr/bin/python


import xml.etree.ElementTree as ET

print 'Parsing XML'

tree = ET.parse('desc2014.xml')

root = tree.getroot()

outfile = open("meshtrees.txt","w")

print 'output ...'

for child in root:

    id = child.find('DescriptorUI').text

    name = child.find('DescriptorName').find('String').text

    tlist = child.find('TreeNumberList')

    if tlist != None:

        treenums = tlist.findall('TreeNumber')

        for treenum in treenums:

            outfile.write('%s\t%s\t%s\n'% ( id, name, treenum.text))

outfile.close()

print 'done'
```

APPENDIX D: DATA DEFINITION LANGUAGE of EACH ENTITY IN CytoCAT

```
CREATE TABLE gene
(
  ID character varying NOT NULL,
  symbol character varying,
  chr_no character varying,
  gene_start integer,
  gene_end integer,
  band character varying,
  strand numeric,
  gc_content numeric,
  gene_db character(50),
  familyid numeric,
  CONSTRAINT "genePK" PRIMARY KEY (id),
  CONSTRAINT "familyFK" FOREIGN KEY (familyid)
      REFERENCES family (familyid) MATCH SIMPLE
      ON UPDATE NO ACTION ON DELETE NO ACTION
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE snp
(  rsid character varying NOT NULL,
  var_allele character(1000) NOT NULL,
  minor_allele character(100),
  minor_allele_freq numeric,
  evidence_status character varying,
  CONSTRAINT "snpPK" PRIMARY KEY (rsid, var_allele)
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE gene_snp
(
  gene_symbol character varying,
  ensid character varying NOT NULL,
  rsid character varying NOT NULL,
  chr_loc numeric NOT NULL,
  CONSTRAINT "genesnpPK" PRIMARY KEY (ensid, rsid, chr_loc)
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE family
(
  familyid numeric NOT NULL,
  family_name character varying,
  parentid numeric,
  CONSTRAINT "familyPK" PRIMARY KEY (familyid)
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE pmp_temp
(
  phenotypeid character varying,
  description character varying,
  meshid character varying,
  term character varying,
  treeid character varying,
CONSTRAINT "pmpPK" PRIMARY KEY (treeid, phenotypeid)
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE mesh
(
  meshid character varying,
  term character varying,
  treeid character varying NOT NULL,
  CONSTRAINT "meshPK" PRIMARY KEY (treeid)
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE phenomesh
(
  phenotypeid character varying NOT NULL,
  hpoid character varying,
  meshid character varying,
  description character varying,
  CONSTRAINT "pmPK" PRIMARY KEY (phenotypeid)
)
WITH (
  OIDS=FALSE
);
```

```
CREATE TABLE phenotype
(
  rsid character varying,
  markerid character varying,
  phenotype character varying,
  pvalue numeric,
  phenotypeid character varying,
  studyid character varying,
  experimentid character varying,
  iscytokine boolean
)
WITH (
  OIDS=FALSE
);

CREATE INDEX markeridx
  ON phenotype
  USING hash
  (markerid COLLATE pg_catalog."default");

CREATE INDEX phenotypeidx
  ON phenotype
  USING hash
  (phenotypeid COLLATE pg_catalog."default");
```

```
CREATE TABLE association
(
  iscytokine boolean,
  gene_symbol character varying,
  ensid character varying,
  rsid character varying,
  markerid character varying,
  phenotypeid character varying,
  phenotype character varying,
  term character varying,
  meshid character varying,
  treeid character varying,
  pvalue numeric,
  experimentid character varying,
  studyid character varying
)
WITH (
  OIDS=FALSE
);
```

APPENDIX E: THE SQL CODE TO CREATE THE *ASSOCIATION* ENTITY

```
CREATE TABLE ASSOCIATION AS (

 SELECT  p.iscytokine,
         gs.gene_symbol,
         gs.ensid,
         p.rsid,
         p.markerid,
         p.phenotypeid,
         p.phenotype,
         t.term,
         t.meshid,
         t.treeid,
         p.pvalue,
         p.experimentid,
         p.studyid

FROM phenotype p

INNER JOIN gene_snp gs ON gs.rsid=p.rsid

INNER JOIN pmp_temp t ON p.phenotypeid=t.phenotypeid  AND
                                   p.phenotype=t.description)
```

APPENDIX F: MESH TERMS OF CYTOKINE GENE RELATED ASSOCIATIONS AND THEIR COUNT.

MeSH Term	Count
Insulin-Secreting Cells	2540
Prostatic Neoplasms	2444
Hemoglobin A, Glycosylated	2328
Body Mass Index	1986
Colitis, Ulcerative	1880
Body Height	1604
Proinsulin	1491
Asthma	1460
Respiratory Function Tests	1053
Insulin	1000
Blood Glucose	977
Arthritis, Rheumatoid	960
Waist-Hip Ratio	946
Amyotrophic Lateral Sclerosis	930
Creutzfeldt-Jakob Syndrome	858
Fibrinogen	832
Alzheimer Disease	768
Bipolar Disorder	768
Cholesterol	764
Drug-Induced Liver Injury	570
Crohn’s Disease	556
Insulin Resistance	535
Parkinson Disease	516
Birth Weight	501
Immunoglobulin E	495
Breast Neoplasms	434
Narcolepsy	422
Schizophrenia	338
Kuru	294
Blood Pressure	221
Epilepsies, Partial	212
Diabetes Mellitus, Type 2	210
Hearing	203
Forced Expiratory Volume	193
Glomerulonephritis, IGA	153
Stroke	144

Macular Degeneration	120
Brain	100
Psoriasis	98
Movement Disorders	85
Diabetic Retinopathy	33
Diabetes Mellitus, Type 1	24
Coronary Artery Disease	24
Immunoglobulin M	12
Alcoholism	6
Hypertension	6
Obesity	6
Dermatitis, Atopic	5
Neuropsychological Tests	5
Myocardial Infarction	4
Tobacco Use Disorder	4
Obesity, Morbid	3
Multiple Sclerosis	3
Lung Neoplasms	3
Celiac Disease	2
Thyroid Hormones	2
Erythrocyte Indices	2
HIV-1	2
Brachial Artery	1
Heart	1
Antidepressive Agents	1
Ankle Brachial Index	1
Skin Pigmentation	1

APPENDIX G : MESH TERMS OF CYTOKINE RECEPTOR GENE RELATED ASSOCIATIONS AND THEIR COUNT.

MeSH Term	Count
Insulin-Secreting Cells	15455
Prostatic Neoplasms	15200
Body Mass Index	11967
Hemoglobin A, Glycosylated	10572
Colitis, Ulcerative	10092
Body Height	9886
Asthma	9052
Proinsulin	8187
Arthritis, Rheumatoid	6572
Waist-Hip Ratio	6226
Amyotrophic Lateral Sclerosis	6130
Insulin	6106
Creutzfeldt-Jakob Syndrome	5904
Blood Glucose	5851
Alzheimer Disease	5325
Respiratory Function Tests	5282
Fibrinogen	4256
Crohn’s Disease	4038
Bipolar Disorder	3915
Cholesterol	3872
Parkinson Disease	3831
Drug-Induced Liver Injury	3285
Birth Weight	3150
Immunoglobulin E	3060
Insulin Resistance	2846
Breast Neoplasms	2514
Narcolepsy	2314
Kuru	2020
Schizophrenia	1844
Diabetes Mellitus, Type 2	1684
Epilepsies, Partial	1475
Stroke	1186
Glomerulonephritis, IGA	1155
Hearing	1119
Forced Expiratory Volume	1100
Blood Pressure	1006

Macular Degeneration	872
Brain	610
Psoriasis	571
Movement Disorders	482
Diabetic Retinopathy	279
Diabetes Mellitus, Type 1	54
Multiple Sclerosis	45
Coronary Artery Disease	36
Obesity	30
C-Reactive Protein	27
Supranuclear Palsy, Progressive	21
Obesity, Morbid	18
Monocytes	18
Alcoholism	16
Liver Cirrhosis, Biliary	15
Morbidity	15
Inflammatory Bowel Diseases	14
Coronary Disease	12
Polycystic Ovary Syndrome	12
Spondylitis, Ankylosing	12
Gallstones	12
Glucose Tolerance Test	12
Colorectal Neoplasms	12
Celiac Disease	10
Leukocyte Count	10
Chemokine CCL2	10
Dermatitis, Atopic	10
Age of Onset	10
Acquired Immunodeficiency Syndrome	8
Neuropsychological Tests	8
Waist Circumference	8
Uric Acid	8
Eosinophils	8
Hypertension	7
Neutrophils	6
Immunoglobulin G	6
Lung Neoplasms	6
Nasopharyngeal Neoplasms	6
Plasminogen Activator Inhibitor 1	6
Cholesterol, HDL	6
Serum Albumin	6

Pancreatic Neoplasms	5
Graves Disease	4
Blood Proteins	4
Behcet Syndrome	4
Cystic Fibrosis	4
Coronary Vessels	4
Hydroxyindoleacetic Acid	4
Carotid Arteries	4
Gout	4
HIV-1	4
Arteriosclerosis	4
Cholesterol, LDL	3
Diabetes, Gestational	3
Skin Pigmentation	3
Subcutaneous Fat	3
Ovary	3
Mucocutaneous Lymph Node Syndrome	3
Plasma	3
Carcinoma, Non-Small-Cell Lung	3
Antipsychotic Agents	3
Body Weight	3
Glioma	3
Brain Neoplasms	3
Sex Hormone-Binding Globulin	3
Tetralogy of Fallot	3
Atrial Fibrillation	2
Ankle Brachial Index	2
Estradiol	2
Erythrocyte Indices	2
Hip	2
Depressive Disorder, Major	2
Bone and Bones	2
Caffeine	2
Blood Cells	2
Atherosclerosis	2
Electrocardiography	2
Attention Deficit Disorder with Hyperactivity	2
Erectile Dysfunction	2
Oleic Acid	2
Fasting	2
Thyroid Hormones	2

Heart Rate	1
Depression	1
Alanine Transaminase	1
Cornea	1
Hair	1
Amphetamines	1
Forced Expiratory Flow Rates	1
Receptors, Leptin	1
Osteitis Deformans	1
Cystatin C	1
Refractive Errors	1
Head	1
Heart Conduction System	1
Glomerular Filtration Rate	1
Pulmonary Disease, Chronic Obstructive	1
Hypothyroidism	1
Intra-Abdominal Fat	1
Panic Disorder	1
Glucose	1
Alopecia Areata	1
Cataract	1
Vitiligo	1