Role of genomic variants in the response to biologics targeting common autoimmune disorders

by

Gordana Lenert, PhD

The thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

Ottawa-Carleton Joint Program in Bioinformatics

Carleton University

Ottawa, Canada

Abstract

Autoimmune diseases (AID) are common chronic inflammatory conditions initiated by the loss of the immunological tolerance to self-antigens. Chronic immune response and uncontrolled inflammation provoke diverse clinical manifestations, causing impairment of various tissues, organs or organ systems. To avoid disability and death, AID must be managed in clinical practice over long periods with complex and closely controlled medication regimens. The anti-tumor necrosis factor biologics (aTNFs) are targeted therapeutic drugs used for AID management. However, in spite of being very successful therapeutics, aTNFs are not able to induce remission in one third of AID phenotypes.

In our research, we investigated genomic variability of AID phenotypes in order to explain unpredictable lack of response to aTNFs. Our hypothesis is that key genetic factors, responsible for the aTNFs unresponsiveness, are positioned at the crossroads between aTNF therapeutic processes that generate remission and pathogenic or disease processes that lead to AID phenotypes expression.

In order to find these key genetic factors at the intersection of the curative and the disease pathways, we combined genomic variation data collected from publicly available curated AID genome wide association studies (AID GWAS) for each disease. Using collected data, we performed prioritization of genes and other genomic structures, defined the key disease pathways and networks, and related the results with the known data by the bioinformatics approaches. We queried the AID results against known data about the aTNFs interventional pathways. Our findings allowed us to infer potential genetic factors and pathways responsible for the aTNF therapeutic effects in AID. A multitude of publicly available bioinformatics tools and databases allowed us to extract the knowledge, analyse it and provide orthogonal evidence for the results. The results support existence of at least two different sets of pathways responsible for AID pathology, most probably reflecting subtypes of AID. Only one set of pathways might be influenced by aTNFs, offering an explanation why aTNF therapy is not always effective.

Additionally, our results narrow down the complex common genomic variability responsible for aTNF unresponsiveness that could be tested in future. If our results are confirmed by functional assays and/or clinical trials, then the lack of response could be predicted ahead of aTNF therapy, leading to better patient selection and improved prognostic outcomes. The extremely high cost of the standard aTNFs therapy can easily cover the cost of genotyping ahead of medication.

Table of Contents

7.3. Supplemental Tables:

7.4. Supplemental Figures:

1. INTRODUCTION

1.1. Motivation

The aim of our research was to investigate genetic variability of common autoimmune/inflammatory diseases (AID) phenotypes, in order to explain the unpredictable unresponsiveness to the anti-tumor necrosis factor biologics (aTNFs) used as the therapy for AID patients.

The anti-tumor necrosis factor biologics (aTNF) is a relatively new class of bioengineered protein drugs. Approved aTNF biologics have been used broadly and successfully for over a decade in clinical practice for treatment of several AIDs. aTNFs have revolutionized AID treatment as the disease prognoses improved significantly, prompting these biologics to a level of the standard of care for AID (Monaco et al., 2014). However, although the aTNF biologics are designed to induce remission by decreasing symptoms and stopping or sustainably reducing disease progression by directly blocking tumor necrosis factor (TNF) in a body, significant portion of patients do not respond to the aTNF for unknown reasons (Taylor 2003; Monaco et al., 2014). The lack of response has emerged as a significant disadvantage for AID patients and a major issue in the aTNF application. The aTNFs unresponsiveness has been a subject of intensive study in the last few years, but no undisputable basis for it has been found, nor have any clear and definite conclusions been made, even using GWAS methodology (Emery 2012; Umićević et al., 2013; Marquez et al., 2014).

In order to explain why aTNFs fail to act in the expected way, we sought to identify the key genes potentially responsible for the aTNFs action in AID genotypes and explore the influence of their variability on the response. Our idea is in accordance with the personalized medicine aspiration and the current pharmacogenomics knowledge progress to tailor drug therapy according to individual's genome, simultaneously achieving the best results and avoiding potential harmful side effects.

Following our hypothesis, we sourced existing publicly available databases and we employed existing bioinformatics software tools to research the relation between AIDs pathogenesis and the intervention action of the aTNF biologics. If the genes responsible for the aTNFs intervention in AID phenotypes could be identified and role of their variants entirely understood, then the variants of the responsible genes in individual AID patients could be tested in advance of the aTNFs treatment. We believe that by testing identified variants, it would be possible to identify non-responders, who then would be assigned to alternative therapies. A selective aTNFs application would prevent inevitable side/adverse effects that stem from the aTNF treatment without being balanced with beneficial AID remission effects. In addition, it might be possible to avoid unnecessary expenses associated with the use of the aTNFs, due to the

very high manufacturing cost and huge demand for the biologics. A similar approach has been implemented in clinical practice for some drugs, where action was found to be highly dependent on variants of the identified genes in the individual genomes (Cavallari et al., 2011). Though still in preliminary phase, this approach represents a foundation of the personalized medicine, where a drug is selected, dosed and applied depending of variants/alleles of the key genes in an intervention pathway of a particular drug.

For the purpose of our research study, we focused only on the AID for which the aTNFs are currently approved as treatment: rheumatoid arthritis (RA) and its juvenile form (JRA), psoriasis (PS), psoriatic arthritis (PsA), ankylosing spondylitis (AS) and inflammatory bowel disease (IBD) including ulcerative colitis (UC) and Crohn disease (CD). We did not include “off label” use of the aTNFs documented in several other AID (uveitis, Bechet), although the convincing data exist about positive outcomes, because the data are not collected in the same manner as for the approved treatments, and the uses are not systematically confirmed to provide unequivocal benefits.

1.2. Autoimmune diseases: burden in society, nature and treatment

Human autoimmune diseases (AID) are heterogeneous group of complex chronic inflammatory disorders commonly present in population. AID are primarily initiated by the loss of immunological tolerance to self-antigens (Janeway 2001). The signature characteristic is the persistent and intermittent chronic immune response to own cells, tissues, organs or organ systems. This immune response reflects disbalance in normally tightly controlled immune system homeostasis. All AID feature adaptive immunity against self-antigens (known or yet not defined antigens), and simultaneously activated innate immunity, both intertwined in a specific manner for each disease entity. The self-directed immune system responses display multiple common and specific characteristics among individual AID phenotypes (Cho and Gregersen 2011). The tissues in which the immune response occurs become inadvertently damaged and dysfunctional over time. Reasons for the persistent immune system disbalance and tolerance break down in AID phenotypes are still partly elusive (Smilek and St Clair 2015).

AID usually start years before a clear clinical presentation, and becomes obvious only after multiple autoimmune-response elements converge into pathological and clinical manifestations. AID have different clinical signs and symptoms, and variable clinical course. Tissue lesions in AID could be multiple, but specific for each disease; they often dictate classification into organ-specific or systemic

AID. Further, AID are classified as seropositive and seronegative depending whether the production of autoantibodies (a humoral part of adaptive immune response type) is associated with the chronic inflammation (Wang et al., 2015).

In addition to very specific characteristics of each disease, dictated by the scope and type of inflammation in organs and tissues, AID phenotypes often share some common epidemiologic, clinical and therapeutical features (Cho and Gregersen 2011; Smilek and St Clair 2015). Recognized shared clinical features of AID phenotypes required similar therapeutic approaches (Cho and Gregersen 2011). This is especially true for AID therapeutic protocols that use the older, mainly empiric drugs with no specific target and with broad effects, discovered quite some time ago on a "trial and error" principle (methotrexate is an example). Invention of such types of drugs is not sought for any more, because better, targeted therapies ask for strictly targeted drugs. However, even now when the targeted therapies are available, many AID still continue to be primarily treated with the non-specific drugs as a first line of a disease control (corticosteroids and chemotherapeutic drugs); the non-targeted, non-specific drugs have opened up a space for development of new more specific, target-based drugs (like TNFs). Oddly enough, the nonspecific medication historically is a standard for comparison and approval of all other newer drugs for the treatment of autoimmune conditions. Target authentication is necessary for the development of new drugs; it requires deeper knowledge of the AID pathogenic mechanisms than the current one. It means that the development of new, targeted drugs, similar to the aTNFs, demands characterization of the molecules capable of inhibiting specific protein–protein or other interactions within the yet unknown pathways that drive disease activity (Song and Buchwald 2015).

As treatable but incurable diseases, AID cause considerable morbidity, affect quality of life and lead to disability. AID are also associated with reduced life expectancy and significant mortality even when threated (Cho and Gregersen 2011; Smilek and St Clair 2015). They are present not only in older adult population, but also in youth and younger adults at the peak of their reproductive and working years. It has been known for a while that some AID constitute a leading cause of death among young and middle-aged women (Walsh and Rau 2000). Autoimmune diseases collectively are ranked within the top 10 causes of death (Jacobson et al., 1997; Wang et al., 2015). In addition, the chronic nature of these diseases places a significant burden on medical care resources, and increases direct and indirect overall economic costs to the society (Jacobson et al., 1997).

The AID as an aggregate affect between 5-10% of individuals of European origin according to some current estimations (Cooper, Bynum and Somers 2003; Marson, Housley and Hafler 2015). The prevalence is similar in populations of Asian, African and Caucasian origin, with few exceptions; however, there is an uneven distribution across other less studied populations, or less studied distinct

diseases (Marson, Housley and Hafler 2015). Incidence rates for individual AID range from less than 5 per 100,000 to more than 500 per 100,000 per year (Cooper and Stroehla 2003). The incidence rate of all AID in the world's population is more than 3% (Cooper, Bynum and Somers 2009), with a greater ratio of females than males. At least 80% of all autoimmune diseases occur in women. However, some AID (Crohn disease, Type 1 Diabetes mellitus) are more prevalent in males. AID affect people of all ages, but disease age of onset tends to be AID specific: for example, diabetes T1D is typical for adolescence, multiple sclerosis (MS) for adulthood. AID incidence and prevalence also differ between geographical regions (Wang et al., 2015). Ethnic differences are similarly documented (Wang et al., 2015). Moreover, rates of specific diseases are noticed to change in real time (Ngo, Steyn and McCombe 2014). A comorbidity of multiple autoimmune conditions are recognized in clinical practice (Cho and Gregersen 2011; Cooper and Stroehla 2003). AID clustering is observed both in individuals and families (Cooper and Stroehla 2003; Cárdenas-Roldán et al., 2013). Comorbidity of the autoimmune diseases is important for interpretation and usage of genomics data (Somers et al., 2009).

Currently, it is widely accepted that AID are complex disorders that evolve from interactions between polygenic endogenous risk factors and environmental trigger factors. However, the relations between environmental and genetic factors and the interactions within genetic factors are still mainly incomprehensible (Smilek and St Clair, 2015). Although certain environmental factors that trigger autoimmunity have been recognized (Pollard 2015), the genetic basis is overwhelmingly accepted as a dominant factor in AID susceptibility and in AID development. It is recognized that AID genetic factors are commonly inherited, but an accumulation of rare acquired somatic mutations could not be completely excluded (Banchereau et al., 2013; Marson, Housley and Hafler 2015). It is important to stress that majority of known genetic factors lack comprehensible relations among themselves and disease pathogenesis (Smilek and St Clair, 2015). Consent is that the molecular pathways ruling pathological (or physiological) events are still very poorly characterized in AID, inevitably leading to poor knowledge of disease characteristics and ways to intervene (Califano et al., 2012).

1.2.1. Genetics of autoimmune diseases

Strong genetic components to autoimmune disease development have been recognized for almost three decades (Goris and Liston 2012). Identification of genetic factors was achieved by traditional epidemiologic, genetic studies including linkage and candidate-gene association, and by positional cloning studies (Altshuler et al., 2008). They confirmed heritable relations with modest effects and an increased allele sharing between affected siblings, validating the shared genetic risk observed in epidemiological studies (Cho and Gregersen, 2011; Marson, Housley and Hafler 2015). Clustering of autoimmune diseases within families commonly appears (Cooper, Bynum and Somers. 2009). Majority

of common AID are characterized by sibling recurrence risk of medium values (Vyse and Todd 1996). The studied AID have higher concordance rates of 25%–50% in monozygotic twins (Bogdanos et al., 2012) versus 2% –12% in dizygotic twins; the results are indicative of genetic involvement in inheritance of diseases (Cooper, Bynum and Somers. 2009).

However, all traditional approaches were insufficient in advancing the insight into the complex genetics of common diseases, as only few genes were identified, insufficient to explain causation or inheritance of AID (Goris and Liston 2012). Better understanding of complex AID genotype architecture and genetic inheritance of AIDs phenotypes was boosted only after introduction of a new methodology of genome wide association studies (GWAS) several years ago. For the first time the GWAS methodology, designed as the case-control model, allowed investigation of genotypes of unrelated people with selected pathophenotypes. It does not require the relatives for detection of genetic inheritance factors, but rather uses huge groups of unrelated individuals with preselected AID phenotypes from various human populations (Goris and Liston 2012).

GWAS methodology advanced investigation of complex AID genomic structures by uncovering polymorphism of numerous genetic loci associated with a single disease. Vast majority of the detected polymorphism loci are common single nucleotide polymorphisms (SNPs), the most abundant source of genetic variation in the human genome (ENCODE Consortium Project 2012). The specific alleles or loci, marked by single nucleotide polymorphisms (SNPs), have been shown to confer small risk increments, defining the predisposition for complex AID phenotypes (Cho and Gregersen, 2011). It is believed that AID, like other common diseases, are “formulated” on common gene alleles according to the “common disease, common variant” hypothesis (Reich and Lander 2001; Altshuler, Daly and Lander 2008; Schork et al., 2009). Because of polygenetic nature and consequently polygenic inheritance, AID might be considered common phenotypes in the same way as complex quantitative traits (such as height) in humans (Manolio and Collins 2009).

The abundance of generated GWAS data became quickly a subject of ongoing investigations on correlations between genotypes and phenotypes. Although GWAS hugely advanced understanding of the complexity of common AID diseases, it will take time to comprehend completely disease mechanisms and even more, to use it as a tool for advancement of future treatments and for prediction of health improvement procedures. The new knowledge generated from advanced genomics is still at the beginning of its application potential (Lucas and Lenardo 2015; Green et al., 2011). Predictive values of AID GWAS genetic results have been intensively studied, but they have not materialized yet as expected to a point that an intervention might be in sight. It is possible that the slow progression of the GWAS-based knowledge about mechanisms of AID has a very good reason: according to the newest

analyses, the heritability of some immune system factors is hard to prove (Brodin et al., 2015). It seems that when it comes to immune system factors and functions, heritability might not go hand in hand with causality. These findings might be interesting if confirmed by other researchers, because the immune system components are directly credited as essential factors for the development of autoimmune diseases.

The GWAS methodology associates well-defined phenotypes with their genotypes; it starts from the preselected clinical phenotypes, and trickles down to identify loci in genomes of selected individuals after scanning them with the biochips. Currently, it is broadly accepted that AID phenotypes are based on genotypes with particular genetic architecture. AID genotypes contain collections of risk factors with modest effects that are commonly present in human populations as allelic variants (usually with frequency ≥1%). However, what is not clear is how risk gene alleles condition extraordinarily complex combination of genetic and potentially epigenetic factors that predispose for autoimmune diseases (Smilek and St Clair 2015; Farh et al., 2015).

As polygenic diseases, the common AID do not have a single highly penetrable genetic mutation as causative factor, but rather a number of multiple causative mutations or variants of very low penetrance that accumulate to gain a signal comparable to single penetrable mutations signals. Some of the single highly penetrable signals are found in human severe immune deficiencies and because immune deficiencies resemble some components of clinical phenotype in AID, they might be considered in AID phenotype-genotype analyses (Fodil et al., 2016). A very few known rare monogenic autoimmune diseases with Mendelian inheritance patterns within families, might be used to help understand genotype-phenotype interactions (Waterfield and Anderson 2010; Lucas and Lenardo 2015). Even in well-studied monogenic diseases, such as sickle cell anemia caused by single point mutations of the hemoglobin beta-chain, the individual pathophenotypes vary significantly, as the patients have different clinical presentations (Steinberg and Adewoye 2006). The phenotypic complexity is also present in all early age onset monogenic autoimmune diseases (Waterfield and Anderson 2010). The examples show that a genotype with a single penetrable mutated gene does not necessarily yield a single pathophenotype. For that reason, it is not unexpected that the common AID diseases also have extraordinarily complex clinical phenotypes, as measured by the disease symptoms, severity and progression (Cho and Gregersen, 2011).

GWAS results have contributed to several conclusions about the genetic architecture of autoimmune diseases. First, the genetic architecture of polygenic AID is highly complex, opposite to the monogenetic Mendelian autoimmune diseases. Second, combined effects of many variants of genetic factors define susceptibility to autoimmune diseases rather than causality. Third, most of detected genetic factors are

common in the general population and each of the genetic factors exerts a small effect on risk for developing AID phenotypes (Hindroff et al., 2009; Cho and Gregersen, 2011). It seems that various different combinations of risk alleles are independently able to generate a high level of disease risk. It is unknown which individual loci are necessary and/or sufficient for the development of AID. Furthermore, it is still unclear whether the disease thresholds exists and what they encompass (Goris and Liston 2012). Whether the genetically susceptible individuals will develop an autoimmune disease is unknown, posing currently an unanswered question “if and when” a specific genotype is sufficient for a disease development, and whether an additional trigger must be taken into account to interact with the susceptible genotype for disease expression (Cho and Gregersen, 2011). Future approaches will be necessary to find rare variants that might be unrecognized by the current GWAS methodology, but suspected to contribute to the overall risk. In addition, it will be essential to understand and integrate analyses of epigenetic mechanisms and eQTL data for better insight into the genetic basis of AID (Farh et al., 2015).

The first genes associated with AID heritability and development were the genes of the human leukocyte antigen (HLA) region characterized more than two decades ago; the extremely polymorphic HLA genes differ in allelic frequency between healthy subjects and patients with AID. However, the nature of influence of the HLA polymorphism on the development of AID is not precisely understood even today, although it is confirmed to be substantial with the numerous AID GWAS. It is still a subject of intensive investigation (Fernando et al., 2008; Goris and Liston 2012; Sollid et al., 2014), solely for one reason: complexity. The HLA system (the human version of the major histocompatibility complex, MHC), is an exceptionally complex system, consisting of a dense collection of two classes of genes: Class I with three major and three minor groups of genes, and Class II with nineteen groups of genes. There is an additional group of genes not strictly of HLA nature and function, with exceptional linkage disequilibrium (Class III), and yet unknown number of HLA pseudogenes. All these HLA gene loci are compactly crammed in the region 6p21.3 of the chromosome 6 in humans; the HLA genes are characterized with high linkage equilibrium (Human HLA Database; HLA Gene Family). There are over thousand common alleles for the major HLA genes and above 13,000 alleles for all HLA genes.

The heritability of HLA genes, though of major significance, cannot solely explain heritability of AIDs, because HLA genes participate in a very narrow functionality of antigen presentation and recognition, and individual sustainability towards foreign invasion (Goris and Liston 2012). The process of antigen presentation and recognition represents a core process of the immune system, but is also important for tolerance development within an individual immune system. Although certain alleles are more observed in specific single AID phenotypes, HLA associations do not have predictive values even when their

heritability has been confirmed (Wang et al., 2015). On the other hand, judging by the disease signs and symptoms, other processes, in addition to the antigen presentation and recognition pathway, must be involved in the development of autoimmune/inflammatory diseases (Cho and Gregersen, 2011).

Before GWAS methodology application, only a few non-HLA risk bearing genes for AID had been identified through candidate gene studies, such as CTLA4 (Ueda et al., 2003) and PTPN22 (Criswell et al., 2005). The GWAS were first to indicate the massive presence of other non-HLA loci associated with AID. It is interesting that well accepted experimental results on genes crucial for autoimmunity in animal models do not overlap with GWAS data. For example, a gene AIRE (autoimmune regulator, a transcription factor) experimentally shown to be a crucial factor for controlling multi-organ autoimmune diseases in experimental animals and human experimental cell models, was only confirmed in RA, and only in patients of Asian ancestry (Terao et al., 2011). Other similar examples have been shown as well. Majority of non-HLA genetic variants are located within non-coding regulatory regions. Recently, to enrich and explain AID GWAS data, fine mapping and QTL expression studies have been performed; they have shown that the variants affect AID by regulating gene expression (Farh et al., 2015).

The AID are also noteworthy diseases from an evolutionary perspective. They have been shown to exhibit a negative effect on the reproductive fitness (Brinkworth and Barreiro 2014). For that reason, AID should have been steadily eliminated from the human population due to the natural selection pressure. There are at least two theoretical explanations why it is not happening (Brinkworth and Barreiro 2014; Raj et al., 2013). The first one stresses that many AIDs occur in the later phases of human life, beyond the reproductive age, due to reduced penetrance of causal genetic mutations. The second, more critical one, points that the persistence of autoimmune conditions might be explained by very essential roles that the implicated immune genes, their products and processes have in immune surveillance and defense. In these cases, the selective pressure is acting to increase host resistance to pathogens and better chances for survival of a population. The genes harboring mutations that qualify as disease AIDs risk factors also might play pivotal advantageous roles in pathogen defense and other life preserving immune functions (e.g. neoplastic surveyance). The same disease risk factors in that case are under very strong natural selection in humans, because they are essential for the survival of the species. Detected risk factors of the immune response are not eliminated from the pool, but propped and enriched (Brinkworth and Barreiro 2014). The results exist to confirm that the risk factors for autoimmunity endure positive pressure due to enhancing certain immune response processes (Raj et al., 2013). Nevertheless, the evolutionary reasons for autoimmune disease persistence in human populations are still not well understood.

1.3. Non-responsiveness to anti-TNF therapy

1.3.1. Tumor necrosis factor alpha

The tumor necrosis factor alpha (TNFα or simply TNF) was discovered virtually simultaneously nearly half a century ago, by two groups of researchers from Yale University and University of California at Irvine (Kolb and Granger 1968; Ruddle and Waksman 1968). TNF is a very potent proinflammatory cytokine synthesized by several cell types not limited to the immune system cells. TNF plays role in immune system activation, differentiation, development, regulation of immune signaling, inflammation, survival or apoptosis and autoimmunity (Aggarwal et al., 2012). It is one of the best-studied cytokines, with known characteristics and roles, and established pathways (Janeway 2001).

TNF and Lymphotoxin (LT α/ β), a molecule very similar to TNF, and their common receptors, all have multiple and important roles in the human immune system. These molecules have been also the subject of study for potential clinical uses (Croft et al., 2013). The essential role of TNF in autoimmunity was revealed after proving its potential to modulate the disease activity in animal models by either increasing or decreasing disease expression (Aggarwal et al., 2012). The first positive effect of TNF blockade was shown in RA in humans, as a proof of concept that its inhibition may ameliorate autoimmunity, revealing TNF as one of the major vulnerable nodes that may be targeted in several AID (Maini et al., 1993).

1.3.2. Anti-tumor necrosis factor biologics

The anti-tumor necrosis factor (anti-TNF) inhibitors (or aTNFs) are approved class of biologic drugs targeting TNF-alpha cytokine. They are engineered monoclonal antibodies or fusion proteins that specifically bind TNF (Willrich et al., 2014). The aTNFs are approved for the treatment of several autoimmune /inflammatory diseases, such as rheumatoid arthritis (Singh et al., 2012), ankylosing spondylitis (Maxwell et al., 2015), psoriatic arthritis (Rodgers et al., 2011), psoriasis (Busard et al., 2014), inflammatory bowel disease (Ben-Horin et al., 2014), Crohn disease, CD (Behm and Bickston 2008) and ulcerative colitis (Lawson et al., 2006). Currently, there are five approved TNF inhibitors: (1) infliximab, a chimaeric IgG anti-human monoclonal (Remicade®); (2) etanercept, a TNFR2 dimeric fusion protein, with an IgG1 Fc (Enbrel®); (3) adalimumab, a fully human monoclonal antibody (mAb) (Humira®); (4) golimumab, a fully human mAb (Simponi®) and (5) certolizumab, a PEGylated Fab fragment (Cimzia®). All five aTNF biologics have been approved as therapeutics/medicinal products by the major regulatory agencies, FDA/EMEA/HC/ (North America and Europe) and TGA/PMDA (Australia and Japan). In addition, several generic aTNF biologics are currently in development, or in the submission process (personal communications). According to the proposed new nomenclature, the five

original aTNFs will be counted as biological originator (bo) aTNF biologics, while biosimilars (bs), after the process of approval, will be named bs aTNF biologics (Smolen et al., 2014).

1.3.3. Therapeutic role of aTNFs

The aTNFs therapeutic outcome in AID is considered revolutionary because it provides overwhelmingly positive results. Introduction of aTNF biologics in late 1990s marked a change in the therapeutic approach, as they become the first successful targeted biologics. Before the discovery of aTNF biologics, treatments for the approved AID (and other autoimmune diseases) consisted only of broad-spectrum immune modulators or traditional DMARDs (DMARD stands for disease-modifying antirheumatic drugs), which all nonspecifically block immune cell functions: non-steroid anti-inflammatory drugs, glucocorticoids, and cytostatics like methotrexate, azathioprine or 6-mercaptopurine and others (Willrich et al., 2015). The advantage of new aTNF biologics, when compared with small molecule drugs, is the result of their dramatically increased target specificity (one that exists for receptors, antibodies), prolonged half-life (proteins) and decreased intrinsic molecular toxicity (naturally occurring molecules) combined (Yin et al., 2015).

The aTNF biologics revised the concept that effective treatments in the autoimmune/inflammatory diseases require broad immune suppression. aTNFs actually challenged the idea that immune pathways are highly redundant and that no single protein targeted therapy might be able to block the inflammation. The aTNFs do work in the complex systems such as AIDs. They proved the concept that upregulation of the immune system, resulting in measurable and damaging chronic inflammation depends on rather fragile, narrow channeled communications between pathways and networks. At least one of communications or pathways integrated into the network of cytokines collapses after neutralization of nodes like TNF. The downregulation of inflammation by aTNFs also supports an idea that a cytokine hierarchy exists and that TNF is most probably a mediator situated near the root of a tree of cytokine networks, negating that the inflammation process is a net of multiple redundant nodes and interactions (Schett et al., 2013). This concept was also supported by the unexpected uniformity of the therapeutic response to TNF inhibition among patients with rheumatoid arthritis, psoriasis, psoriatic arthritis, Crohn disease, ulcerative colitis, ankylosing spondylitis, juvenile arthritis and other less prevalent autoimmune diseases. However, there are exceptions such as multiple sclerosis and SLE, which worsened upon TNF inhibition (Zhu et al., 2010; Probert et al., 2000). Their pathogenesis might be different from aTNF sensitive AID, although they are classified as the same group of diseases, exhibit similar symptoms or signs, and, even more, are treated in a similar fashion.

Given the profound therapeutic potential of TNF inhibition, it has been suggested that different chronic inflammatory diseases (AID and others), may share common pathophysiology based on the notion that aTNFs are able to disrupt the easily broken obligatory inflammation network. It has been also suggested that even more targeted therapies might follow in future, based on better understanding of aTNFs action in AID (Schett et al., 2013).

1.3.4. Efficacy and safety of aTNFs

Current approach to AID treatment focuses on induction of remission and requires chronic management that extremely rarely, if ever, results in a complete cure. aTNFs are given over a long period of time and although very effective, their tapering, interruption and cessation are subjects of current investigations with no consensus. Recent evaluation of the dose reduction, discontinuation or disease activity guided dose tapering of the aTNF agents produced limited conclusions, because of heterogeneity of study data (Malottki et al., 2011).

Unwanted consequences of the aTNFs medication are not trivial, since aTNFs exert characteristics of the immunosuppressive drug class. aTNF adverse events (AE) could be detrimental to effective immune responses and might diminish natural ability of the immune system to fight infections and to resist and control malignant processes. Among AE that might surface during aTNFs therapy, cases of tuberculosis (TB) and lymphomas as the most serious AE occurred in practice (Nanau and Neuman 2014). The reactivation of tuberculosis (TB) has been a recognized risk of aTNF biologics therapy from the beginning (Xie et al., 2014). The current studies are inconclusive regarding the TB incidence rate (British Thoracic Committee 2005; Nanau and Neuman 2014), and not sensitive enough to assess the risk of reactivation of latent TB infection vs. de novo inoculation (Sivamani et al., 2013). The longer-term safety data for the malignancy risk varied across studies, but none reached statistical significance (Singh et al., 2012; Smolen et al., 2013). Also encouraging is the finding that the immune response to the tested vaccines (pneumococcal and influenza vaccines) is not reduced in patients during aTNF therapy (Hua et al., 2014). In addition, aTNFs potentially may cause rewiring of an individual immune system leading to secondary autoimmunity (Singh et al., 2012; Smolen et al., 2013).The use of aTNF biologic drugs has been linked with the paradoxical development of systemic and organ specific autoimmune processes, such as sarcoidosis (a granulomatous disease) and other diseases (Korta et al., 2015). The effects of TNF inhibitors on cardiovascular disease are potentially multifaceted, because these drugs may promote heart failure, but on the other side may also improve risk factors for atherosclerosis (Nguyen and Wu 2014).

All five aTNF agents are effective in the induction and maintenance of remission; there is no evidence of clinical superiority among them in clinical trials. Nevertheless, aTNF biologics differ in the way they individually are dissolved and administered, by their serum peak and trough levels, *in vivo* complexes that they can form, and incidence and timing of rare AE (Michaud et al., 2014; Malottki et al., 2011).

The aTNF biologics are not always effective for all patients, and even when effective initially, might lose effectiveness over time. Even in clinical trials, initial failures of aTNF induction therapy occurred in up to 40% of patients. Like many drugs, aTNFs are less effective in daily clinical practice than in clinical trials, due to the less controlled environment (drug delivery, adherence, other medication, etc.). A proportion of patients with AID that do not respond adequately to treatment with anti-TNF drugs in clinical practice might be even higher (Kamal et al., 2006). The aTNF treatment guidances state that lack of response warrants re-evaluation, as it is considered most probable an intrinsic failure of the treatment (Saag et al., 2008; Smolen et al., 2013). Switching from one aTNF agent to another aTNF after first-line treatment failure may not be a cost-effective treatment strategy (Kamal et al., 2006). However, switching, while controversial, occurs frequently because of patients' comfort and experience (or lack of it) with the aTNF agents. Occasionally, a second agent from the same class may be effective due to individual differences in bioavailability and immunogenicity. The practice is widespread, as the 2006 US survey indicated that >94% of rheumatologists had switched patients to another TNF inhibitor, because of lack of efficacy or intolerability of the first agent (Kamal et al., 2006). When more than one TNF inhibitor provides inadequate responses or AE, then switching to a completely different class of biological agents with another mode of action (and there are at least four approved non-TNF targeted biologics) may provide a more effective option (Emery 2012). In our opinion, all these data indicate that aTNFs have very narrow if not a single path of action, because if one aTNF agent does not work in a patient, other aTNF agents do not work as well.

Secondary loss of aTNFs response is also a common problem with incidence ranging between 23% and 46% at one year after anti-TNF initiation. Immunogenicity of aTNF drugs is one of the mechanisms behind this type of treatment failure (Garcês et al., 2013). Though anti-drug antibodies (ADA) are often presumed as underlying mechanisms for failure of the treatment with a biologic, it is hard to confirm the suspicion by currently available tests. Measurements of aTNF ADA are complicated and expensive and do not always provide an adequate answer that can be used for clinical decisions (van Schouwenburg et al., 2013). In some patients, the formation of anti-TNF ADA might be transient (van Schouwenburg et al., 2013).

1.4. GWAS, variant database, tools, and analyses; genotype-phenotype relationship

1.4.1. Importance of genome wide association studies

Genome wide association studies (GWAS) is an advanced bioinformatics methodology that relates human phenotypes to human genotypes. GWAS application associates intrinsic, inborn known common variability of human genomes with preselected phenotypes expressed as complex human traits or common diseases (Hindorff et al., 2009; Schprk et al., 2009). Testing is performed on hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) at a time, in the genomes of thousands of individuals. The GWAS methodology enables researchers to analyse polygenic disease/trait inheritance, for which the classical genetic methodology did not and could not provide an answer.

By late 1990s, the existing genetic mapping of human complex diseases became unpractical and insufficient, because of the very high number of genes suspected to be involved in common diseases. The used human genetic maps, with several hundreds of loci marking human chromosome regions, became obsolete for all complex diseases and traits, except for rare single gene mutation based Mendelian diseases. The genetic markers, labeling DNA polymorphisms detected by positional cloning, were too scarce for studying inheritance of complex diseases and traits in humans. Unsuccessful were the attempts to identify genes for complex diseases with pedigree analyses, as the pedigree studies were only able to identify single highly penetrant genes influencing a range of monogenic diseases/traits, which include a variety of inborn errors of monogenetic genetic diseases such as cystic fibrosis, Duchenne muscular dystrophy, or Huntington disease (Borecki and Province 2008). However, most disease genes were completely unsuspected based on existing knowledge at the time, which made “candidate” genes approach obsolete. In addition, large informative pedigrees, which this type of analyses needs, are very difficult to find and to recruit (Borecki and Province 2008). These analyses are curbed by many limitations: clinical, ethical, and intrinsic, such as locus heterogeneity, incomplete penetrance, and variable expression (Altshuler, Daly and Lander 2008). In the Mendelian diseases, variations in a single gene are both necessary and sufficient to cause a disease. Level of necessity and sufficiency for disease genes in complex diseases and traits are unknown parameters. The known Mendelian disease and genes related to them have been documented extensively in the Mendelian Inheritance in Man or OMIM NCBI database (McKusick, OMIM database).

Alternative approach for disease genes discovery emerged from population genetics and genomics, by means of comparisons of frequencies of genetic variants among affected and unaffected individuals (Risch and Merikangas 1996). Further reasoning led to the so-called “common disease–common variant” (CD-CV) hypothesis: the brilliant idea proposing that common polymorphisms (defined as

having a minor allele frequency of >1%) might contribute to susceptibility to common diseases (Risch and Merikangas 1996; Lander 1996; Schork et al., 2009).

The GWAS methodology offered a solution for polygenic variant identification in complex common diseases unrestrained by any presumptive hypothesis (Altshuler, Daly and Lander 2008). No previous knowledge of potential gene "candidates" is required for GWAS analyses, making it a method of choice for genetic background in common diseases, since common diseases have been considered for long to arise from the interactions between greatly unknown polygenic risk factors and poorly understood environmental factors (Selmi, Lu and Humble 2012). However, the heritability of common diseases has been firmly established, derived from phenotypic concordance in twins and first-degree family members as well as familial clustering (Czyz et al., 2012).

Before introduction of GWAS, the common diseases heritability in general unrelated population could not be studied. GWAS design follows case-control study design with very large groups of unrelated individuals. Unrelated subjects with common diseases are much easier to recruit; their recruitment location became less important than their willingness and consent to participate in genetics studies. Both are regulated and more demanding than sample collection, mainly because of effortlessness to gather DNA samples for GWAS (Peppercorn 2012). It is important to stress that the CD-CV hypothesis did not assume that all common disease are conditioned only by common causal mutations. Common variants exist in human population because of the nature of human population formation and they are used to detected loci for detailed study of common human diseases or traits. Contrary to the simple understanding that the CD-CV hypothesis means only common variants may cause common diseases, a full spectrum of alleles has been expected always (Lander 1996; Altshuler, Daly and Lander 2008; Schork et al., 2009).

GWAS have been highlighted as a major breakthrough in health research over the last decade (Altshuler, Daly and Lander 2008; Zerhouni and Nabel 2008), not only because of the potential to explain genetic origin and heritability of common diseases, but because the new studies have capacity to provide new feasible therapeutic and prevention approaches. GWAS methodology has currently a central role in human genetics revolution as it has facilitated substantial massive data accumulation for further studies and the advancement of knowledge in medicine and biology (Zerhouni and Nabel 2008).

1.4.2. GWAS methodology

GWAS enable advancement towards the fundamental goal to completely catalog genetic polymorphisms of the human genome that cause phenotypic variations and allow characterization of molecular

mechanisms by which polymorphisms exert their effects (Manolio, Brooks and Collins 2008; Manolio and Collins 2009). GWAS methodology is considered currently the most effective way of mining for variants (Manolio and Collins 2009), because it is purely discovery driven and unbiased towards any particular hypothesis. It became a standard tool for detection of variants, which are deemed inheritable risk factors, responsible for developing of complex diseases. Not only that numerous GWAS examined the human genome to detect risk variants, but they also revealed genetic architecture of human diseases and polygenic traits in healthy people. GWAS genome scanning potential have also facilitated new discoveries in genetic recombination (Hinch et al., 2011; Hindorff, Gillanders and Manolio 2011) and better understanding of natural selection and evolution (Novembre et al., 2008).

GWAS methodology is based on finding small minuscule variations in a form of nucleotide switches that are naturally occurring in human genomes throughout the evolution of the species. Usually any two human genomes differ in one DNA base pair in every thousand base pairs, but the number could be higher (Lander 2011). The single nucleotide polymorphisms (SNPs) are the most common type of DNA variations in the human genome. A typical GWAS may genotype significant number of SNPs, starting from 300K to more than million SNPs. It is typically testing the genomes of thousands of individual subjects who are preselected based of their expressed phenotypes (diseases/traits) and assigned into case or control cohorts (Hinds et al., 2005; Altshuler, Daly and Lander 2008). Finding SNPs association is performed by scanning of genome DNA for preselected markers manufactured on biochips, using highly sophisticated detection equipment and procedures. Common genetic variants generally have very small effects that require large sample sizes for detection (Visscher et al., 2012). GWAS methodology must meet stringent levels of statistical significance (e.g. $p < 5 \times 10^{-8}$), as it has been justified because of multiple hypothesis testing. The association results must be replicated ideally in a separate sample of cases and controls, to minimize the possibility of assay artefacts (Pe'er et al., 2008).

The following projects and innovations allowed the GWAS instituting, as they became the core around which genome-wide association study methodology was built:

1.4.2.1. Human Genome Project

The successful completion of the Human Genome Project (HGP) was an unprecedented scientific achievement; it was the absolute milestone in biology and definitely a turning point in medicine. The HGP has provided an essential foundation for sequencing and analysis of additional human genomes and has laid the foundation for all other discoveries about the human genome. The HGP became an invaluable resource in the search for genes that cause human diseases (monogenic and polygenic). Interestingly enough, the project even started as a motivation to find genes responsible for diabetes,

many years before publications of the first drafts of the human genome (Venter et al., 2001; Lander et al., 2001).

The HGP established the consensus human reference sequence. It found that the approximately 3.3 billion nucleotides constitute the human genome. Particularly, the human genome seems to encode only 20,000-25,000 protein-coding genes, less than originally estimated (Lander et al., 2001). A very small part of the human genome is determined to be protein-coding sequences, as the total length covered by the coding exons is approximately 34 Mb or ~1.2% of the genome; the untranslated regions of the transcripts are estimated to cover another ~21 Mb or ~0.7% of the genome (Lander et al., 2001). The analysis and establishment of the reference sequence was complicated by the presence of polymorphism in the DNA samples, because it was hard to distinguish whether differences between sequence clones reflected errors or polymorphism (Lander et al., 2001).

1.4.2.2. HapMap Project

While the reference sequence constructed by the Human Genome Project is informative about the massive bases in DNA sequence that are invariant across individuals, it did not capture the natural variation of the human genome. The Haplotype Map of the Human Genome Project (HapMap) was founded to focus on DNA sequence differences and to determine common pattern of variation across the human genome (The International HapMap Consortium 2004, 2005). The HapMap Project was launched in 2002, and was a natural extension of the Human Genome Project (The International HapMap Consortium 2003; Lander 2011). The HapMap Project has helped understanding human diversity. It also has helped detect positive natural selection across the human genome (Sabeti et al., and The International HapMap Consortium 2007).

A major motivation for the HapMap was to catalogue and compare variations in various human populations, find their frequencies and associations and make it publicly available for medical research. The idea about helping medical research was based on notion that inherited genetic variation has an important role in the pathogenesis of disease. Interestingly enough, the seminal paper on HapMap Project cited that “the root causes of common human diseases remain largely unknown, preventative measures are generally inadequate, and available treatments are seldom curative” and further stated those reasons as a motivation for the project work (The International HapMap Consortium 2005). The comprehension of genetic background of human diseases has demanded examination of genomes of individuals and comparison of all genetic differences between controls and affected groups. The comparison between groups ultimately could have been accomplished by complete sequencing of the genomes of individuals under study. However, more practically, an alternative comparison between

existing common variability in human groups, by which the humans differ, is used for finding variation difference in GWAS analysis (Hinds et al., 2005). Today, after facing collection of enormous population variation data and near exhausted analyses of all existing data for some diseases, the search for disease causing variants is back to sequencing methods: a whole-genome sequencing that enables characterization of almost all variants in an individual or, a whole exome sequencing (WES) that provides all variants in protein coding genes.

The HapMap project helped to guide the design and prioritization of SNP genotyping assays for GWAS methodology by documenting the substantial recombination hotspots, structures of linkage disequilibrium and low haplotype diversity linking correlations of SNPs with many of their neighbors (The International HapMap Consortium 2005; Manolio, Brooks and Collins 2008). This feature has been used for finding the relevant SNPs in common diseases.

The HapMap project has achieved a high-density haplotype mapping, first starting with smaller number of population representatives (few representative groups), searching variations in blocks of 500 kb, and later extending it to 100-kilobase regions and larger reference panel (eleven different populations instead of the original four). Advanced phases of the HapMap project encompass variants not only restricted to common SNPs (MAF ≥ 5%), but also low-frequency (0.5 %< MAF<5%) variants that account for the vast majority of the heterozygosity in each sample, and a large number of rare (0.05 %< MAF <0.5%) and private (singletons and MAF <0.05%) variants (The International HapMap Consortium 2007, 2010). At that point, the project was also able to detect copy number variations. With the HapMap extensions or phases, a very slight improvement happened for finding common SNPs, but there was greater improvement in genotyping for low frequency and rare SNPs (The International HapMap 3 Consortium 2010). HapMap 3 catalogued both allele frequencies and linkage disequilibrium (LD) between new variants. The vast majority of common SNPs strongly correlate to one or more nearby proxies, which lead to the conclusion that 500,000 SNPs can provide excellent power to test >90% of common SNP variation in out-of-Africa populations, with roughly twice that number required in African populations (The International HapMap Consortium 2005).

The HapMap project originally genotyped one million SNPs, but updates over years increased that number to more than 10 million common DNA variants, primarily SNPs, although still in a limited set of DNA samples (The International HapMap 3 Consortium 2010). The vast majority of common variants either is represented in dbSNP database, or is in tight correlation to other SNPs that are represented in dbSNP database (The International HapMap 3 Consortium 2010).

The HapMap 3 data were obtained with the Affymetrix Human SNP array 6.0 biochip (interrogating 1,852,600 genomic sites) and the Illumina Human1M-single beadchip (interrogating 1,199,187 genomic sites), that were initially applied to 1,486 and 1,284 samples, respectively (The International HapMap 3 Consortium 2010). The HapMap project data constantly gets updates with improvements in sequencing technology, especially for low-frequency variation. We used HapMap 3 SNP data for our research as the most current.

1.4.2.3. Linkage Disequilibrium

Linkage disequilibrium (LD) is defined as a non-random correlation between two loci (two different SNPs or other markers) or allelic association between two alleles. The LD phenomenon is known to exist in the human genome (Gabriel et al., 2002), but vary within human populations. It signifies haplotype maintenance through the evolution. The physical arrangement of SNP alleles along a chromosome is called an "haplotype" (Olivier 2003), and it became an essential characteristics exploited in the GWAS methodology. Properties of LD in the human genome have been studied in detail by HapMap project with the aim to understand both the causes of LD and its application to disease research (Olivier 2003; The International HapMap Consortium 2005).

LD is a consequence of common ancestry and recombination pattern in human genome and is a measure of coinheritance of human haplotypes. LD is defined by parameters providing information about how often two or more loci or alleles are inherited together (D) and how related to each other they are depending on their frequency in the population (r2). In the case of high LD, both measurements are close to 1 (on a scale ranging between 0-1).

The principle of LD at the population level allowed for GWAS methodology, as GWASs are based upon LD. When a sufficient number of SNP markers or tags is available for testing in a genetic study (like it is in GWAS), any common variant, even if it was not assayed directly, should display significant LD with neighboring marker SNPs. A set of evenly spaced SNPs at high resolution across the genome should permit whole-genome association studies. Furthermore, it has become clear that the extent of LD and haplotypes in the human genome is not simply a function of distance between SNPs. The HapMap project found that the LD varies markedly on a scale of 1–100 kb; the LD not only varies in length, but is also often discontinuous and does not proportionally decline with a distance (The International HapMap Consortium 2005). The data reveal that SNPs are typically perfectly correlated to several nearby SNPs, and partially correlated to many others (The International HapMap 3 Consortium 2010). Considering only common SNPs (the original target of study for the HapMap Project) in the population of European ancestry or CEU, one in five SNPs has 20 or more perfect proxies (with perfect LD of 1), and three in

five SNPs have five or more perfect proxies. In contrast, one in five has no perfect proxies (The International HapMap Consortium 2005; The International HapMap Consortium 2007).

1.4.2.4. 1000 Genomes Project

After several years of GWAS, it became clear that a more in-depth look at variation, including rare variation, was necessary to explain additional disease heritability and relationships between phenotype and genotype. The new project, 1000 Genomes Project (1KGP) was designed to provide detailed description of common human genetic variation by applying whole-genome sequencing of healthy people and their progeny belonging to different human populations (The 1000 Genomes Project Consortium 2010). Variations have been obtained by direct sequence comparisons, instead of exploring variations using allele frequencies, by applying a combination of low-coverage whole-genome sequencing, and deep exome sequencing. The 1KGP sequenced 2500 individuals and gathered information on variants in different populations. It provided accurate haplotype information on all forms of human DNA polymorphism in multiple human populations (The 1000 Genomes Project Consortium 2012, 2015). Advances in DNA sequencing technology have enabled the sequencing of individual genomes for the 1000 G project (Bentley et al., 2008).

Coming after massive GWAS implementation, the 1KGP boosted more detailed and more accurate GWAS data analyses, overcoming limitations of the HapMap project as the only reference for GWAS. The first results from the 1000 Genomes project (the pilot project) were reported in 2010 (The 1000 Genomes Project Consortium 2010). Results of the 1KGP have provided researchers with a population scale map of rare variants to complement and enrich existing knowledge of common variants gained from the HapMap project.

The data from the 1KGP have provided the location, allele frequency, and local haplotype structure of approximately 15 million SNPs (The 1000 Genomes Project Consortium 2012). With the advancement in number of individuals within additional new populations, the 1KGP was able to uncover 84.7 million single nucleotide polymorphisms (SNPs) in addition to other variations of human genomes (insertions and deletions, copy numbers, to name a few), that are far less common in humans (The 1000 Genomes Project Consortium. 2015). The 1KGP, in addition to discovering common variants (MAF ≥ 5%), estimated at 8 million, has enabled fine mapping of loci for both low frequency (0.5% <MAF <5%) estimated at 12 million, and rare (scarcer) variants (MAF < 0.5%) and private variants, estimated at over 64 million. Regarding population specificities, most common variants are shared across the world, and rarer variants are typically restricted to closely related populations. A major chunk of SNP variability (86% of variants) is restricted to a single continental group. Up to one third of all SNPs for European

ancestry population (CEU) detected by the 1KGP were novel SNPs (The 1000 Genomes Project Consortium 2010, 2012, 2015).

The project found that the public databases were less complete for coding SNPs at low frequencies, especially for the lower frequency SNPs outside the exons. Of special interest was the finding that detected variations were not evenly distributed across the genome: certain regions, such as the HLA, show very high rates of variation, while others do not (important for common autoimmune/inflammatory diseases).

Although it is difficult to conclude accurately the number of functional SNPs for each individual genome, collectively there were around 10,000 synonymous SNPs and 11,000 nonsynonymous SNPs as reviewed by the project. Approximately one in twenty of nonsynonymous SNPs was found to be damaging SNPs (loss of function (LOF) and damaging function SNPs), which means that an individual might have between from 5 to 250 damaging SNPs per exome, as estimated by extrapolation of the results, apparently depending on the used methodology (The 1000 Genomes Project Consortium 2012, 2015). Overall, an individual genome might have between 3-4 million SNPs, but only 40,000 to 200,000 SNPs (1-4%) of them are not common SNPs and have a frequency <0.5%. The 1KGP found that a typical genome contains 149–182 sites with protein truncating variants, 10,000 to 12,000 sites with peptide-sequence-altering variants, and 459,000 to 565,000 variant sites overlapping known regulatory regions (untranslated regions (UTRs), promoters, insulators, enhancers, and transcription factor binding sites). It also observed around 2,000 variants per genome associated with complex traits found through genome-wide association studies (GWAS Catalog) and 24–30 variants per genome implicated in rare disease through the Clinical Variation database (ClinVar) (The 1000 Genomes Project Consortium 2015).

Comparison of the two data sets of detected SNPs from the HapMap project and from the 1000 Genomes project revealed that approximately 72% of HapMap SNPs were also found in 1KGP data. After filtering out HapMap variants with a MAF of <5% (separately for each population), 99% of HapMap SNPs were found in 1KGP data (Buchanan et al., 2012). However, the rare variants differ more between projects than the common variants. Although the number of 1KGP variants increased several times, when the recalculation of variants are performed for a particular GWAS, the number of imputed common and intermediate frequency variants increased by 7%, whereas the number of rare variants increased by >50%. These differences in the detected polymorphism might affect SNP queries or imputation used in GWAS analyses. The number of novel variants is constantly increasing and many believe that the 1KGP potentially has overshadowed the utility of HapMap (Buchanan et al.; 2012).

Both the HapMap and 1KGP projects have contributed to identifying linkage disequilibrium patterns and consequently the ability to choose tag SNPs for further GWAS studies. Based on completely different technologies, they also contributed to studying genomic structure, recombination rates, and mutation rates.

In our study, we used data based on the human genome LD results and SNPs data obtained from both HapMap3 project and 1KGP. All GWAS SNPs studied in our research have been confirmed by the 1KGP. For the all GWAS SNPs there was a difference between two sources only in very few SNPs.

1.4.2.5. The Encyclopedia of DNA Elements (ENCODE) project

Once the human genome sequence is established, the next step was the identification and annotation of functional DNA elements in the human genome. The Encyclopedia of DNA Elements (ENCODE) project has been established with the aim to identify and catalog all genomic sequences according to their functions (ENCODE Project Consortium 2004, 2007, 2012). The ENCODE pilot first was tasked with annotating predefined 1% of the human genome or 30Mb with functional elements (ENCODE Project Consortium 2004, 2007). The experience and methodology from the pilot project have been consequently extended to rigorous analyses of elements with biological information of the entire genome. ENCODE project is an ongoing collaborative effort that is using different technologies and approaches to identify functional elements (ENCODE Project Consortium 2007, 2012).

The ENCODE has produced high-resolution reproducible maps of DNA segments with biochemical signatures associated with diverse molecular functions (Lander 2011; Ernst et al., 2011; Gerstein et al., 2012). ENCODE project define functional elements as discrete sequence elements that confer biological function. They might be segments that have reproducible biochemical functions (such as binding sites or other chromatin distinctive structures) or serve as templates for defined products (proteins or non-coding RNAs). They are mapped as genes, transcripts, regulatory regions, chromatin sites and methylation patterns. ENCODE project changes the definition of genes to include regions of genome outside of the well-studied protein-coding regions and not only narrow continuous space around protein-coding regions (Rodriguez-Fontenla et al., 2014). Generated detailed maps came from few dozen human cell lines (ENCODE Project Consortium 2011). The ENCODE data in particular enabled assigning various potential biochemical functions for up to 80% of the genome (Kellis et al., 2014). In contrast to evolutionary and genetic evidence, biochemical data offer clues about both the molecular function of DNA elements and the cell types in which they are active (Kellis et al., 2014). However, the jury is still out regarding the definition of what constitutes biochemical functionality of the huge part of genome, when it is known that only up to 10% of the human genome is filtered under pressure of the natural

selection and conservation (Brunet and Doolittle 2014). The findings are not widely accepted because the authors are not predicting how many functional elements such genome might claim (Kellis et al., 2014).

Why is ENCODE important for GWAS? ENCODE data might provide a starting point to study human disease in addition to differentiation and development of particular cells (Ernst et al., 2011; Gerstein et al., 2012; Kellis et al., 2014), as they can be linked with SNP risk factors detected in GWAS. The ENCODE maps might help interpretation of GWAS signals, connecting SNPs with potential mechanisms of gene regulation (Kellis et al., 2014). ENCODE data for a particular region of interest can be retrieved from HaploReg database using search with preset options of LD, type of population, source etc., and from RegulomeDB using its own algorithms for prediction of regulatory significance of elements. Other tools for visualization are available from UCSC Genome Browser (Kent et al., 2002). The browser can be used to show functional features in the region of interest (ENCODE Project Consortium 2010, 2011).

In our study, we used ENCODE data for our research, since the ENCODE project enabled searching for a SNP potential change of a functional element, especially the AID SNPs that reside in the non-genic regions with no clear known functions.

1.4.2.6. Genotyping biochip

Essentially critical for the implementation of GWAS have been the advancements in biochip manufacture technology. Standard chips are structured to detect tag SNPs relatively evenly distributed along the whole genome. The number of SNP tags has constantly increased over time, making standard chips able to provide better resolution suitable for high-density mapping (Eberle et al., 2007). However, in most cases high resolution was not enough to distinguish an association signal due to a direct risk SNP from an indirect association signal due to a LD effect. Standard genotyping chips currently used for GWAS are not well suited for either picking up the remaining common variants that are not yet discovered, or for identifying rare variants, due to the limitations of the technology and because of occurrence of perfect LD (Cortes and Brown 2011).

Other genotyping limitations of the chips in use have been also recognized. The majority of genotyping chips are designed for use in European populations. They might be less informative for other ethnic groups, especially admixtures, particularly if the SNPs are not shared between populations. Another weakness is that many rare variants have yet to be identified and are not represented on the chips (Cortes and Brown 2011; Lee et al., 2014).

Some newer microarray-based genotyping technologies permit flexibility in choosing the scope and density of SNP markers. The development of custom genotyping chips such as the Immunochip designed for immunogenetics studies, the Metabochip designed for studying metabolic diseases, or a CD chip for cardiovascular disease, has further advanced GWAS research.

Initiated by the Wellcome Trust Case-Control Consortium, the Immunochip was designed by leading investigators covering all of the major autoimmune/ inflammatory diseases (Cortes and Brown 2011). These specific biochips have better resolution in the regions of interest, but on the other side, they might be biased, as they limit a GWAS to the narrower predetermined loci, ones that were selected based on previous knowledge and assumptions coming from it. They potentially may leave undetected loci that have been previously unsuspected loci, undetected loci, loci across the whole genome and loci not related to the previous knowledge (Cortes and Brown 2011). However, at loci with established disease association, the Immunochip contains all known SNPs in the dbSNP database, as well as from the 1000 Genomes project 2010 release (The 1000 Genomes Project Consortium 2010).

Taking into consideration the latest 1KGP results, the current GWAS chips obviously do not identify rare variants very well. Although the companies are now racing to increase rare variant coverage on genotyping chips, it might not be the easiest task, and plain whole genome sequencing (WGS), or even whole exome sequencing (WES), are becoming more popular as a more accurate way to find rare variants. Nevertheless, very high-density chips such as the 5 million SNP chips are existing products in the Illumina pipeline (Cortes and Brown 2011).

With the advancement in human genome knowledge and technology, it is necessary to understand complexity of findings and take into consideration a design of studies and especially interpretation of rare variants results (Lee et al., 2014).

1.4.2.7. Other advancements contributing to GWAS implementation

Additional important advancements that made GWAS possible were technological advancements in computation and informatics, which enabled data storage in various structured biological databases, web-based tools for storing and sharing data and management of unprecedented amount of data for further search and processing. The advancement in construction of databases has been combined with the capability of providing almost endless innovative ways to search for and analyze data (Merelli et al., 2014). These innovative approaches have mined data to provide new results, previously not accessible by any experimental method. The new protocols have become real experiments in the virtual spaces (Agarwal et al., 2014).

Change of social sharing attitudes, mandatory open access policy for scientific publications, overall public educational advancements, easier information accessibility and exchange, and other changes in social behaviors, all have contributed to a certain extend to broaden ideas of human genomic research. All instances allowed for recruitment of individuals eager (and curious) to provide their DNA specimens necessary for human genome research and GWAS applications (Turner et al., 2011).

Essential to the GWAS implementation is data sharing: fact is that the GWAS research has been critically dependent on data and ideas sharing among various professional and research groups. Exceptional collaborations and huge consortia formations have been necessitated by almost unsurmountable number of human subjects needed for the studies and by complexity in management of generated data.

As GWAS emerged as a powerful unbiased tool for discovering the genetic basis of complex human diseases or traits, and number of SNP markers genotyped by GWAS has dramatically increased, the GWAS statistical design and analysis has been addressed by multiple researchers groups. The goal of statistical improvement has been to identify as many significant features in the genome as possible, while incurring a relatively low proportion of false positive and false negative results (Lee et al., 2014).

The testing of a large number of potential variants in parallel, requires a stringent threshold for significance in order to limit false positives (type 1 errors), meaning that discovery of variants responsible for small effects requires very large sample sizes (Storey and Tibshirani 2003). The large number of correlated markers demanded the multiple hypotheses testing correction (Han et al., 2009). Failure to adjust for multiple testing appropriately might produce excessive false positives or overlook true positive signals. Multiple testing is a challenging issue in genetic association studies not only because of large numbers of markers, but also because many of them exhibit linkage disequilibrium (LD) and are not independent entities (Storey and Tibshirani 2003). The approaches that assume independence may be highly conservative, including Bonferroni correction that is often thought to overestimates the statistical penalty of performing many correlated tests (Gao et al., 2008).

The current concept is to control the false discovery rate (FDR), or the expected proportion of false positives among the rejected hypotheses (Benjamini and Hochberg 1995). The FDR (also called the Bayesian false discovery rate) is a sensible measure of the balance between the number of true positives and false positives in many genomewide studies (Storey and Tibshirani 2003). From a practical viewpoint, the procedure is simple: input the p-values for the genes into an FDR software, get the output of the corresponding q-values (Storey and Tibshirani 2003), and then declare an event significant, if its q-value is less than or equal to 0.05. This supposedly ensures the FDR to be controlled at 5 % level.

Novel approaches have been also recently explored, in order to increase the heritability estimation from detected GWAS SNPs and to enhance the detection of newer, rare SNPs that might contribute to explaining heritability without increasing the size of study groups (Yang et al., 2011; Liley and Wallace 2015). Some novel statistical approaches have been proposed recently, but the clear benefits are still unknown; the aim is to extract more SNPs without redoing new studies or increasing the number of subjects, or simply to provide more accurate results (Lee et al., 2014; Lin and Lee 2015). A statistical approach plays a crucial role when it comes to making decisions on cut-off values and exclusion/inclusion of GWAS detected SNPs.

1.4.3. GWAS database

As the interpretation of GWAS findings in the context of the disease biology remains challenging, it is essential to make all GWAS data as accessible as possible for the broader scientific community for further analyses. The most complete public database of GWAS data is the NHGRI catalog of GWAS (Welter et al., 2014). The NHGRI Catalog is manually curated and maintained by a professional staff, because the data extraction is an expert activity. The GWAS Catalog data are solely extracted from single studies or meta-analyses, typically reported in peer-reviewed publications (Welter et al., 2014). Usually GWAS data are deposited in the dbGaP as well, where users must apply for access to individual-level genotype and phenotype data and comply with a data access agreement (Mailman et al., 2007). Deposited GWAS in the NHGRI catalog must conform to the eligibility criteria, including number of SNPs assayed and non-targeted study design (Hindorff, Gillanders and Manolio 2009). The Catalog has been constantly under development. A recent improvement is to structure the trait annotations using ontology principles. All data in the GWAS Catalog will be integrated in an ontology format that can be processed by an ontology reasoner and queried. The schema ontology has been designed to model the key concepts represented in the Catalog: trait, SNP, study, gene, chromosome and relationships that link these concepts (Welter et al., 2014). The availability of mappings between GWAS traits and ontology terms facilitates the integration of the GWAS Catalog with other resources.

GWAS Catalog data can be accessed in few ways. GWAS Catalog data are available through commonly used data portals such as PheGenI tool among others (Mailman et al., 2007). Data are also accessible via the web interface hosted at the NHGRI.

A major challenge for the Catalog future development is the increasing complexity of studies that include ethnicity extraction, search interface to support ontology-driven querying, gene interactions, and accommodation of next generation sequencing technology, as well as improving links to other resources.

1.5. Genotype–phenotype relationship in AID

1.5.1. From SNPs and genes to molecular networks and pathways

Genome wide association studies (GWAS) have collected valuable data that provide useful insights into the genetic heritability of many common diseases, including the AID. However, not surprisingly, the plethora of GWAS data has remained unused to its full potential, because the biological mechanisms underlying genotype-phenotype relationships in common diseases remain only partially explained, even for the most studied diseases like T1D (Schork et al., 2009; Carter, Hofree and Ideker 2013). Very few mechanisms by which the common diseases occur are known, leaving majority of GWAS findings outside our useful incorporated knowledge.

Counterintuitively, the GWAS data collected to explain heritability of common diseases, also may help us comprehend underlying biological mechanisms, by finding missing links and putting forward predictions for experimental assessment (Goldstein 2009; Califano et al., 2012). The rationale is that, if a SNP detected by a GWAS is a risk factor for common disease development, than a gene deviated (changed or influenced) by the same SNP is the gene that participates in common disease pathogenesis.

From the very beginning of GWAS data usage, quite odd and unexpected relationships have been discovered for some common diseases. Macular degeneration has been extremely significantly linked to H factor (CFH gene) of the complement system (Fritsche et al., 2014), creating a connection that was never suspected before. IL28B (renamed IFNL4) gene and its pseudogene condition the chronic viral hepatitis C infection and clearance (Suppiah et al., 2009), but this gene was even unknown before and never discovered to be linked to immune system. However, some crucial genes have been omitted for unknown reasons, such as C4 and C3 genes from human complement system not being picked by numerous schizophrenia GWAS until recently, when the genomic variations of nearly 65,000 people were analyzed (Daly et al., 2016), linking them to this common disease.

In addition, because of the knowledge gaps, the overwhelming wealth of common disease genomic data has neither materialized into the development of new therapeutics, not even for the most studied common diseases like diabetes mellitus; nor it helped explain why certain drugs do not help in treatments of some patients (Goldstein 2009; Califano et al., 2012). This application of GWAS data is now in the focus of many pharmaceutical companies, which do not share genomic data with the broad scientific community. Consequently, this lack of knowledge hampers to great extent a goal of

personalized medicine to assess and use patient's individual genotype to guide clinical care (Fernald et al., 2011).

In common diseases, where the heritability is based on many genes, expressed disease phenotypes are reflection of various pathobiological processes that interact within a complex network; a disease phenotype also progresses over time as gene expression and epigenetic control change (Barabási et al., 2011). The genetic variants (overwhelming majority being SNPs) condition disruptions/modifications of coordinated numerous biological processes within cells and tissues, and establish their transformations into pathobiological processes that result in common diseases phenotypes (Barabási et al., 2011).

In the near future, it is likely that a catalog of virtually all human genomic variations will be produced. However, it is still daunting question how this amazing knowledge might be translated into knowledge about disease mechanisms (Vidal, Cusick and Barabási 2011). There are still major unsurmountable problems regarding how to model human genetic variation impact on formation of pathogenic pathways and networks (Vidal, Cusick and Barabási 2011).

The phenotypic impact of gene variant modifications, including deficiency of any specific gene product caused by SNPs, is not determined solely by the function of a mutated gene, but also by the functions of components with which the gene and/or its product interact (Goldstein DB. 2009). In other words, the impact of any mutation in a gene (SNPs are mutations) is determined by its product network context, making its impact contingent on a very complex modified network that might be changed at multiple points simultaneously (Califano et al., 2012). Even when the interactants are not changed per se, the impact of an original altered gene has multiple outcomes, usually hard to predict and comprehend in advance (Barabási, Gulbahce and Loscalzo 2011).

The potential complexity of the human interactome network is daunting: with 20,000 plus protein-coding genes, encoding yet undefined number of splice variants and post-translationally modified forms of proteins, significant but not yet defined number of functional RNA molecules, and with about a thousand metabolites, the distinct cellular components that serve as the components (nodes) of the interactome easily exceed one hundred thousand (Barabási, Gulbahce and Loscalzo 2011). The number of functionally relevant interactions between the components of such a huge network is expected to be much larger and remains unknown (Venkatesan K et al., 2008).

For the exploration of the complex interplay between human interactome and human diseases, it might be first desirable to comprehend current molecular pathways and phenotypic network maps. The

protein-protein interaction maps have been constructed with great effort and organized into comprehensive databases, such as MINT, IntAct, BioGRID, HPRD, which are still considered largely incomplete. Metabolic pathways are probably the best identified part of interactome, with KEGG metabolic map collection being the most complete database of human pathways. Mapping of human regulatory networks is in its infancy. RNA networks, encompassing RNA-RNA and RNA-DNA interactions, are still work in progress. However, for some types of RNAs, like microRNAs, interaction data are organized into several comprehensive microRNAs databases with predicted miRNA targets (Barabási, Gulbahce and Loscalzo 2011).

When the human interactome was analyzed as a network by applying network theory rules (Barabási and Albert 1999), the presence of highly connected hubs was determined (Jeong H et al., 2001). These hubs hold the whole network together and have specific roles in the human interactome. They encompass highly connected proteins that are classified into “party” hubs (that function inside specific cellular processes) and “date” hubs (that link together different specific cellular processes). Evidence from model organisms indicates that hub proteins tend to be encoded by essential genes (Fraser et al., 2002), and that genes encoding hubs are older and evolve more slowly than genes encoding non-hub proteins (Saeed and Deane 2006).

When the position of disease genes in relation to hubs was researched, it was found that essential genes, but not the disease genes, encode hubs (Jeong et al., 2001). From an evolutionary perspective, it is understandable that mutations that disrupt hubs have difficulty being transmitted in the human populations as they negatively influence huge number of biological processes and cannot be kept in the gene pool because of their lethality. Disease genes obviously thrive in the human population, meaning they are transferable because they do not influence survival. Only mutations that functionally impair peripheral genes can persist, forming a pool of mutations conditioning heritable diseases most of which appear in adulthood (Barabási, Gulbahce and Loscalzo 2011). Peripheral genes do not show a tendency to encode hubs and tend to be tissue-specific (Goh et al., 2007). That means most of the common disease genes could be classified as peripheral genes. At the same time, it was discovered that peripheral genes have fewer interactions and might be involved in fewer pathways (Bossi and Lehner 2009). These tissue specific and more recently evolved proteins make fewer interactions than core proteins and participate in fewer pathways. The older a protein is, the more interactions it makes (Bossi and Lehner 2009).

It was also found that newer, peripheral and most likely disease genes interact with hub genes. They interact with hub genes even more often than with each other. Most tissue-specific proteins bind to universally expressed proteins, and function by recruiting or modifying core cellular processes into

tissue-specific biological processes. On the other hand, most 'housekeeping' proteins that are expressed in all cells also make highly tissue-specific protein interactions; possibly all 'housekeeping' proteins actually have important tissue-specific molecular interactions (Bossi and Lehner 2009).

The hubs making essential genes are proposed to be involved in tissue specific processes as different splicing variants adapted for specific tissue processes. Understanding tissue-specificity is often instrumental to understanding complex diseases, because particular interactions may occur in one type of tissue with participation of expressed, specific and core proteins, while they do not occur in another type of tissue. The same set of genes is present, but they remain inactive in other tissues; when they are not expressed simultaneously, there are no interactions, although theoretically they may occur. In other words, when proteins qualified to interact are not expressed and brought into physical proximity at the same time, they do not necessarily interact in particular tissues even if they belong to hubs (Bossi and Lehner 2009; GTEx Consortium 2015).

This concept of tissue specific and disease genes among them, interacting with modified older core 'housekeeping' genes might help find specific pathways for disease genes and modulate them with applicable drugs in an easier way than projected. It is well recognized that, while the genetic variants that cause hereditary diseases are global (detected in germline) and present in all cells across the entire human body, the diseases are often only manifested in specific tissues. The mechanism for this tissue selectivity and vulnerability at the same time is unknown (Barshir et al., 2014).

1.5.2. Basic concepts of biological networks and pathways

Because the biological processes are accomplished via proteins, RNAs and intermediary products, a potential way to explore connections between GWAS results and phenotypes is to identify pathways or networks where these genes execute their roles, and connect their functions with a specific disease. Depending of their expressed variants, proteins behave differently in pathways, or in networks, because protein variants influence the speed and size (quantification) of biological processes. Most likely, the same is true for nucleic acids, where a change in their sequences (such as SNPs) might influence their functional roles (van der Sijde et al., 2014).

Given the complex genetic architecture and synergistic effects among genes, the holistic effect of a network or a pathway is expected to have a larger effect than the sum of the individual effects of each gene in the structure (Khatri, Sirota and Butte 2012). This approach also cuts complexity, increases the power of analyses, and, importantly, allows for easier biological verification (Khatri, Sirota and Butte 2012). Extrapolation from SNPs to pathways or networks via affected gene products (usually, but not

necessarily proteins), also overcomes the problem of SNP variation in different human populations with the same phenotypic outcome of a specific disease. Various SNPs present in human population groups within the same region might have the same functional consequences. Associated pathways might be more consistently replicated across ethnic groups, compared to the SNP-level association for the same type of disease (Schaid et al., 2012).

Consequently, using the knowledge of predefined canonical pathways (majority are experimentally discovered), the findings of disease-associated pathways can be complementary to the single SNP/single gene analysis for better understanding of the molecular mechanisms in diseases (Khatri, Sirota and Butte 2012). Combining data from GWAS, these SNP-pathway approaches can assess whether groups of genes with related functions are jointly associated with a disease of interest and help generate specific hypotheses for follow-up experimental studies (Carter, Hofree and Ideker 2013; Khatri, Sirota and Butte 2012).

Networks are usually defined as collections of nodes that are joined together by edges; edges could be with or without direction. Biological network models are built from experimental data, but also from queries of the literature databases and public databases of molecular interactions (Carter, Hofree and Ideker 2013). Nodes in the networks might represent various types of molecules, not only proteins, with edges representing relations between nodes not limited to direct molecular interactions (physical or chemical); edges might be defined as other types of relations: co-expression, co-localization, influence on gene regulation, co-regulation by a third molecule, feedback regulations, etc. (Szklarczyk et al., 2015; Carter, Hofree and Ideker 2013).

1.5.2.1. Biological pathways

A biological pathway represents a set of molecules connected with series of actions among molecules that results in an end-product. Its definition has evolved to describe a pathway as a set of interacting genes (or their products) that together performs a very specific biological function. A biological pathway is the strictest definition of a gene set. Relations between members, almost exclusively proteins, are restricted to biochemical reactions or direct physical interactions that build or resolve molecular complexes. Biological pathways are vector-driven structures towards an essential, specific end-point and are finely balanced by quantities and location of every component and reaction, including its regulation or regulation of adjacent pathways (positive or negative feedback). Biological pathways are usually derived from experimental work over time, but with development of bioinformatics software and databases, they might be constructed by imputation and then tested experimentally.

More recently, four types of pathways have been proposed in attempt to describe the heterogeneity of currently available pathways: molecular, cellular, disease, and intervention pathways (Mooney et al., 2014; Ramanan et al., 2012). Molecular pathways characterize biochemical actions on a molecule or compound; cellular pathways model the regulation of more global cellular processes. It was noted, however, that disease and intervention pathways may simply be collections of genes previously associated with a phenotype, or sometimes combination of basic pathways (molecular or cellular), rather than being based on knowledge of precise biological mechanisms (Ramanan et al., 2012). In addition, concerns about the hierarchical nature of biological pathways have been raised, meaning that some pathways can represent subsets of larger pathway modules. Evidently, the pathways may differ by size (20-200 members), complexity and data sources (Mooney et al., 2014). These differences can have consequences for the interpretation, confidence, and comparability of results from gene set analyses (Khatri, Sirota and Butte 2012).

Pathways are usually much smaller structures than networks, with fewer members and relations than networks, because they are structured usually on experimental data and annotated by a few groups of experts (Barabási, Gulbahce and Loscalzo 2011). Pathways might be defined as a very specific narrowly defined type of a directional network that has a start and an end.

In an ideal case, GWAS candidate loci can be linked to a phenotype using canonical pathways (Carter, Hofree and Ideker 2013; Goldstein 2009). However, in most cases, candidate genes implicated or imputed by GWAS studies are not well characterized and their products have not been included in any known canonical signaling/metabolic pathways. Additionally, a good portion of candidate genes is not qualified beyond its sequence, with no function assigned to it and no annotations. Some candidate genes even have no defined orthologues to provide additional information about their potential function.

Even when they exist, however, canonical pathways are likely to be incomplete and even inaccurate (Carter, Hofree and Ideker 2013; Califano et al., 2012). Some estimates go as far as to claim that many pathways are not known currently and are yet to be discovered (Carter, Hofree and Ideker 2013; Califano et al., 2012), as current data shows that more than 90% of known proteins are not located in any pathway. Systematic screens of the proteome suggest that canonical pathways capture only a fraction of the true protein-protein interactions that occur within the cell (Guruharsha et al., 2012). Furthermore, biological pathways are often treated in studies that connect pathways with GWAS data as a set of related genes, which jointly perform a biological function, ignoring the specific and directional relationship among the members. This approach of treating a pathway as a network creates partial and vague pathways and not directional pathways, driven in a directional mode with an end-product.

The linking of GWAS SNP harboring genes within canonical pathways is not only a labor-intensive task; the uncertainty of its execution and outcome is relatively high.

Adding to the problem of defining pathways that hypothetically underlie diseases under GWAS, is the fact that the associated GWAS SNP loci are not always tied with any protein coding genes, even after analyzing for LD (van der Sijde et al., 2014; Hindorff, Gillanders and Manolio 2011). Most of GWAS-implicated risk variants reside outside of protein coding genes. These GWAS SNPs influence non-protein-coding genes or not yet identified genes, as they appear to cluster in stretches of regulatory DNA sequences or regions of DNA with no known formations (Maurano et al., 2012). It has been recently suggested that the majority of the genome is involved in biochemical and regulatory activities of cells, not just the 1%-1.5% of the genome responsible for encoding proteins (ENCODE Project Consortium 2012 ref18). This discovery is augmenting the complexity of currently existing pathways with additional potentially huge number of unknown interactions (Dunham et al., 2012). New classes of molecule such as microRNAs and lincRNAs are increasingly implicated in regulating the activity of protein coding genes (Fernald et al., 2011), but they are not integrated into biochemical or signaling pathways, which at this point are only consisting of protein members. Non-coding genes are currently implemented only in a very few canonical pathways and only in cancers (according to the data available in pathway databases such as KEGG, BioSystems, and NCI Pathways). In cancers, however, the somatic human variability is often different from the heritable variability (SNPs in germline) discovered in common diseases. Hence, even when integrated into a few biological pathways, these non-coding gene variants are not applicable to common diseases. Furthermore, the pathways might be wired in a different way in neoplastic cells compared to non-cancerous cells (ENCODE Project Consortium 2012).

Non-coding genetic alterations, even those affecting non-coding RNA (ncRNA) sequences, are suspected to mediate phenotypic effects primarily by altering the abundance of proteins in the cell and thus perturbing Protein-Protein Interaction (PPI) networks through stoichiometric effects (Esteller 2011; Carter, Hofree and Ideker 2013). Functions of these so-called regulatory SNPs could be very complex and elusive, and involve gene expression regulation through the effect on RNA splicing, transcription factor binding, DNA methylation and miRNA recruitment (Knight 2014; Huang 2015). Using ENCODE up to 80% of all previously reported associations data have some kind of functional annotations and mainly represent regulatory SNPs (Schaub et al., 2012; Ward and Kellis 2012). For a majority of associations, the SNP whose functional role is most strongly supported by ENCODE data is a SNP in linkage disequilibrium with the reported SNP, not the genotyped SNP reported in the association study (Schaub et al., 2012). However, reported functionalities currently are not incorporated into some structured biological processes, pathways or networks. In addition to regulatory non-coding SNPs, recent advances

in the theoretical and experimental methods used to study DNA packaging within cells, elucidated the biological function and pathways to which SNPs located within gene deserts can contribute (Schierding et al., 2014).

An investigation of the tissue-specific effects of genetic variants on gene expression uncovered surprisingly complex relationships (Fu et al., 2012; Carter, Hofree and Ideker 2013). Unknown complex relationships between different molecules at this stage of the knowledge are suggesting that network models may be better suited than pathways for dissecting phenotypic consequences of non-coding variants (Ward and Kellis 2012; Kellis et al., 2014).

Canonical pathways are almost always represented as linear chains of events in order to better visualize relationships between molecules and to be able to make predictions for experimental tests in a manageable number of experiments. However, cell regulation is anything but linear; it is instead determined by complex multivariate interactions not amenable to visual interpretation (Califano et al., 2012). Pathways are oversimplified representation of biological processes and are not always able to explain processes (Kellis et al., 2014).

Biological pathways are collected and integrated in databases by initiatives like KEGG (Goto S. et al., 1997), Reactome (Joshi-Tope G et al., 2005) or the user-curated WikiPathways (Pico AR et al., 2008), and many other pathway databases.

1.5.2.2. Biological networks

In contrast to canonical pathways, biological networks represent a view of all connectivity among a large number of genes (Carter, Hofree and Ideker 2013). A biological network typically defines a partnership among molecules, regardless of actual molecular functions its members may perform (Sun 2012; Khatri, Sirota and Butte 2012). Unlike pathways, networks do not explicitly describe a specific biological function or process carried out in a specific biological context. Networks simply aim to describe biological relationships (observed or predicted interactions) between multiple genes or gene products. In general, the heterogeneous evidence is used to build publicly available PPI networks, and they are not centered on biological process (Mooney et al., 2014). Networks are composed mostly of proteins as gene products (e.g., PPI network) and define not only their physical interactions, but also co-expression, mutual regulation, co-localization, co-mentioning in the same article (after text mining) or biochemical reactions in experimental data etc., as shown in STRING and other databases (Szklarczyk et al., 2015).

Biological networks represent a collection of all known relationships in the knowledge space and are less stable than the well-curated canonical pathways. They might easily be expanded to contain protein-DNA interactions, or protein-DNA-ncRNA interactions, because these networks are not limited to a functional outcome. The understanding of these biological relationships is steadily growing, hence causing biological networks to expand and change (Sun 2012). However, various biases, including those introduced by measurement technologies, might be carried when the networks are constructed (Khatri, Sirota and Butte 2012).

Genetic variations such as SNPs that are influencing the abundance or activity of individual molecules, consequently affect the interactions and networks in which those molecules participate (Carter, Hofree and Ideker 2013). It has been proposed that to understand genotype-phenotype relationships it will be necessary to quantify the effects of mutations on molecular networks. These changes can have a spectrum of consequences, ranging from completely abrogating protein activity to having no effect at all, to even occasionally enhancing it (Wang et al., 2011; Carter, Hofree and Ideker 2013). A variety of computational strategies has been developed to predict the functional consequences of mutations at the protein level (Jordan et al., 2010; Sunyaev 2012).

Network- and pathway-based methods have been developed to boost the power of the candidate genes identified by GWAS and to bridge the gap between genetic variants and biological phenotypes (Sun 2012). According to some evaluations, these data have enabled a good part of the human genome to get some assignment of biochemical functions, in particular outside of the well-studied protein-coding regions (ENCODE Project Consortium 2012; Kellis et al., 2014).

1.6. Biological Ontologies

The term “ontology” has a long history in the philosophy since Aristotle. In the philosophical sense, ontologies are systems of categories that account for a particular way of seeing the world. The goal of building ontology is to construct an exhaustive and definitive classification of all entities (Gruber 2009).

In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose to enable the modeling of knowledge about some domain (Gruber 2009). Thomas Gruber defined “an ontology as a specification of a conceptualization” and “intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals”, and further simplified that “ontologies are what they do: artefacts to help people and their programs communicate, collaborate and

coordinate" (Gruber 1993). Ontologies simplify unification among programs and serve in the role of data exchange and analyses. They are constructed in a specific domain for the purpose of enabling domain knowledge to be shared and reused. The term 'ontology' is also frequently used to refer simply to controlled terminologies. Controlled terminology defines the common vocabulary in which shared knowledge is represented (Gruber 1993).

New biotechnological discoveries from the last two decades have generated an unprecedented amount of information. At the same time, informatics and information technology became able to manage huge amounts of data and provide tools for their analyses. The amount of information in both bioscience and biomedicine has exponentially increased, especially with the acquired "omics" knowledge (Rubin at al., 2006). Researchers often face problems with extracting the data they need from the large numbers of data sources that are available. It has become a central problem to find ways to readily integrate information and data from biomedical and biological domains, and make it available for searching, sharing, reusing, analyzing, making new extrapolations and generating hypotheses. Difficulties to cope with data have necessitated repositories with the capability to be well maintained and useful (Lazakidou 2010).

According to the latest NAR Database issue, currently there are over 1,500 databases listed in the Molecular Biology Database Collection. All data are typically stored in publicly available databases on the web and often paired with informatics tools for their mining and analyses. In the bio-molecular/biomedical domain, experts have been annotating biological entities with controlled vocabularies, introducing terminologies that have allowed common sharing and understanding of biological entities and processes they establish and partake (Lazakidou 2010).

In order to compute the huge factual scientific knowledge, the used semantics has to be precisely defined and standardized (Gruber 1993). To achieve this aim, terminologies and ontologies are used as instruments (Rubin et al., 2006). There are differences between the two formations. Terminologies use standardized and precisely assigned controlled vocabularies as a collection of descriptors for all entities in a domain, and principally represent simple collections of names in a domain, but do not provide relationships between entities. On the other hand, ontologies contain controlled terminologies (controlled vocabularies) plus essential components of the semantic network that encodes relationships between each term of the vocabulary. Ontologies represent networks of controlled vocabulary terms (nodes) and relationships (edges) between the terms. Relationships are mapped in a hierarchal order so that the whole structure becomes an organized "construction". Ontologies represent the powerful instruments for knowledge representation, as they enable formalized knowledge to be structured as required for

computational processes, so that a machine or a person can explore the web of data (Malladi et al., 2015).

The National Center for Biomedical Ontology (NCBO) became the leading scientific organization for bringing semantic technology to biomedicine (Musen et al., 2012). The NCBO's mandate has been to build a comprehensive library of biomedical ontologies and to create tools and methods allowing researchers to use the ontologies. As a result, the BioPortal resource was created as an open source repository of biomedical ontologies that stores and provides access via the web. BioPortal provides investigators, clinicians, and developers with 'one-stop shopping' to programmatically access biomedical ontologies. BioPortal also supports linkage to data from other biomedical resources. This ontology library is curated and updated by the administrators of BioPortal and the bioinformatics community, which has access to annotations, and mapping of entities (Noy et al., 2009).

The best-known ontology in the biomedical domain is probably the International Classification of Diseases (ICD-11). Main bioscience related ontologies are compiled in the Open Biomedical Ontologies, or OBO Foundry (http://www.obofoundry.org/), and are collaborative projects involving developers of science-based ontologies (Smith et al., 2007). The ontology projects are ruled by the established principles and methodology for ontology development with the aim to construct and maintain a group of connected reference ontologies in the bioscience/biomedical domain (Smith et al., 2007). Many more very useful ontologies have been constructed (Dumontier et al., 2014; Shah, Cole and Musen 2012). Among OBO Foundry ontologies, probably the best-known and most widely used ontology is Gene Ontology.

1.6.1. Gene Ontology

The GO project is a community-based ontology that creates evidence-supported annotations to describe biological roles of individual genomic products (e.g. genes, proteins, ncRNAs complexes, etc.) by classifying them using ontology principles. The Gene Ontology development started more than a decade ago focusing on the need to integrate knowledge about genes and their products with consistent annotation (Gene Ontology Consortium 2000). Gene Ontology is managed by the Gene Ontology Consortium (GOC) with more than two hundred collaborators. The GOC has been involved in various projects and collaborations with the goal of expanding and improving the representation of biology (Gene Ontology Consortium 2015). The GO is a dynamic ontology open to new additions and modifications that are inevitably introduced as scientific knowledge expands (Shah, Cole and Musen 2012).

The GO project has developed three ontologies that describe gene products in terms of their associated characteristics connected with biological processes (BP) they participate in, molecular functions (MF) they execute and cellular components (CC) as sites of action. The GO provides ontology-type organized and controlled vocabulary information for the properties of genes, thus sorting them into three domains. For each of these domains there are distinct terms, which describe gene or gene product properties. The biological process domain provides sets of molecular events with a defined beginning and end. The molecular function domain defines the activities of gene /gene products at the molecular level. The cellular components domain describes the parts of cellular or extracellular environment as a location attached to a gene/gene product.

In addition to maintaining ontology, the project itself has different aspects such as developing the ontology and its tools and maintaining annotations of genes and gene products. GO resources include gene product annotations in the form of comprehensive statements about what gene products do, usually in a species-neutral fashion (Gene Ontology Consortium 2000). The completeness of annotations derived from biomedical literature is uneven for each entity and across the BP, MF and CC ontologies for the same entities. Controlled vocabularies have been structured and maintained and the GO can be queried at different levels. The GO structure, relations, and terminology are modified regularly by GO editors, making it a complex and perpetually changing dynamic organization (Gene Ontology Consortium 2015).

One of the main uses of GO is to perform enrichment analysis on gene sets (Schaid et al., 2012). The Enrichment analysis tool is one of the software tools incorporated into GO, and is simple to use, relatively straightforward and yields easily interpretable information. GO Enrichment analysis software tool provides a mechanism to determine statistically significant functional subgroups within gene groups. Using this software tool is a simple way to gain insight into potential biological significance of a gene set under study. This software tool can determine whether GO terms associated with the particular biological process, molecular function, or cellular component are over-represented in the group of genes deemed significant by the statistical analysis, which allows one to gain insight into the potential biological significance of a gene set under study (Shah, Cole and Musen 2012).

Most recently, the GO descriptors have been tied to biological pathways and diseases. They could be searched using GO tools. Enrichment for BP, MF and CC terms based only on experimental data and not only on available literature (text) is also currently available in the beta version (since May 15th 2015), in addition to complete data from three ontologies (Mi et al., 2013). As proven, the data mining of the GO can be used for the discovery of new biological associations and even new drug targets.

We used the GO for extracting knowledge about hundreds of SNP bearing/related genes discovered by AID GWAS, including BP, MF, CC, pathway and disease domains. The collected GO descriptors may clarify the potential of these genes to influence AID development and indicate their underlying pathology.

1.7. Research Approach

As stated before in the Abstract and Introduction sections of the thesis, the objective of this study is to find the key factors that influence a therapeutic action of the aTNF biologics in AID disease phenotypes (or lack of it).

aTNFs have proven therapeutic (curative) effect on the AIDs, indicating that the same interventional processes may work in several AIDs with different clinical manifestations. Different clinical manifestations stem from disease processes targeting different tissues, but AIDs also have some shared mechanisms as a group of diseases (autoimmune/inflammatory disease have been mostly treated by the similar drugs even before aTNF introduction). If aTNFs, as very specifically targeted drugs, all have positive response in majority of AID patients and are able to induce remission, then some parts of the common pathogenic AID mechanisms must be engaged by aTNFs, resulting in remission. Shared mechanisms between AID disease processes and aTNF interventional processes must overlap to certain extent, and they are responsible for remission upon aTNF therapy. We wanted to find out what particular common set of interactions is responsible for disease modulation. However, if the same set of interactions permissible for aTNF action does not exist in some AID diseased individuals, because it is functionally modified, an AID phenotype cannot be influenced by aTNF drugs, resulting in some degree of non-response.

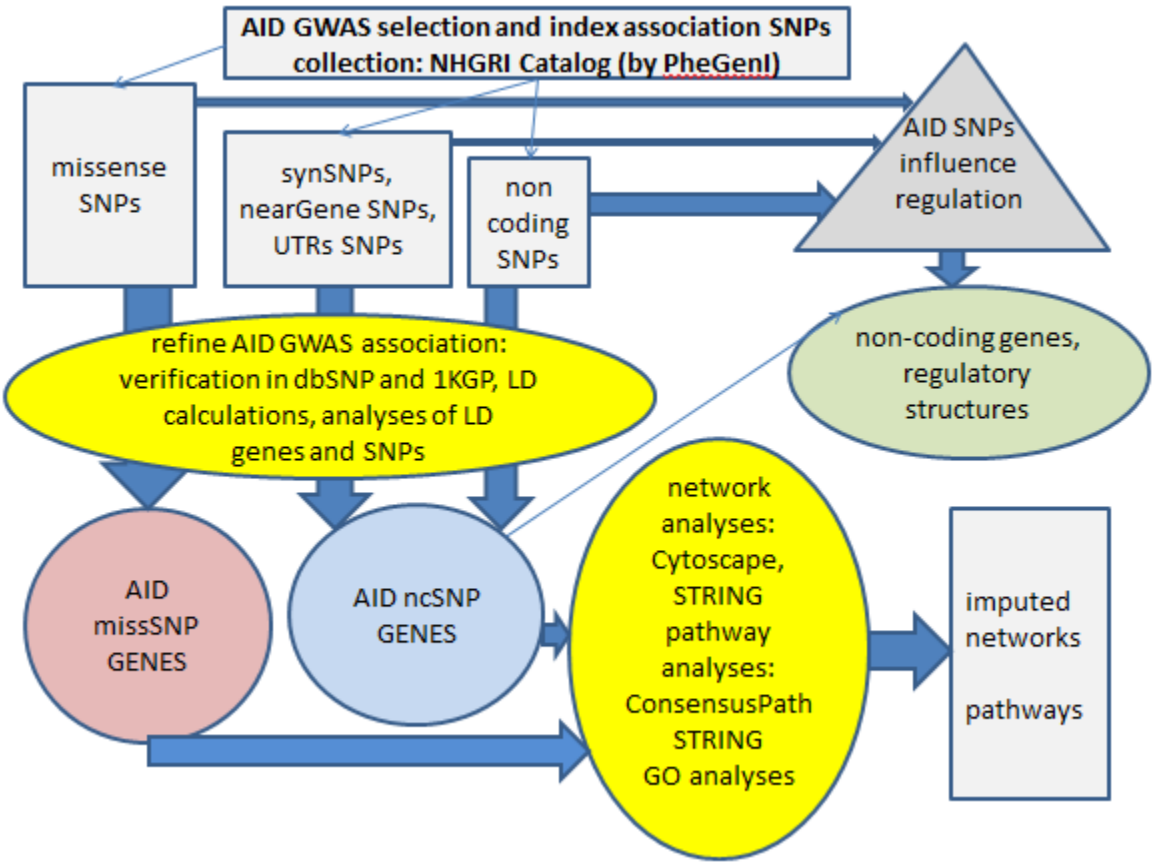
We investigated the intersection between TNF and AID pathways (between intervention and disease pathways) in order to uncover a common set of interactions that may be responsible for modulating aTNF effect and consequently for variable responses to the aTNF therapy due to variability of the constituents. Our unique research approach integrates the knowledge of the aTNF biologics intervention and the knowledge of AID genotypes-phenotype relationships solely cropped from publicly available GWAS data and other publicly accessible bioinformatics databases. Combining the AID GWAS data with the relevant information from other databases and by using advanced bioinformatics software tools for retrieving and analyzing the data, we aimed to unveil potential network elements, pathways and genes responsible for AID pathogenesis and aTNF response.

The null hypothesis is that if no such network can be constructed, then there may be other mechanisms responsible for unresponsiveness of AID patients to aTNFs, mechanisms that are not necessarily linked to the natural inheritable human genetic SNP variability.

Our research plan consisted of the several steps that were executed in a sequential order. The first step was to find SNPs associated with the AID disease phenotypes from published GWASs, and then select and retrieve association data. The following step was to analyze the retrieved data using available bioinformatics software tools in order to detect SNP locations and SNP relations with the known genomic structures in the human genome (genomic structured that currently have annotations). In next few steps, depending of the type of SNPs and genes, we defined which genes harbor AID SNPs and explored impact of AID SNPs on characteristics and functions of these genes and gene products. Following the step of SNP-gene prioritization, we further explored potential role of the AID SNP genes/proteins in the known biological processes by several approaches. Using bioinformatics software tools, we completed canonical pathway enrichment analyses, network analyses, GO terms enrichment analysis, and disease enrichment analysis, all of which helped us determine functional properties of AID SNP genes. We then used the acquired knowledge about functional bioprocesses in AID for comparison with known aTNFs intervention actions as drugs. By combining the data from AID SNP functional exploration with the TNF action data, we have determined common pathways and genes we consider the key elements for therapeutic effect of the aTNFs. They are key elements because they determine curative actions of the aTNFs in AID.

2. MATERIAL AND METHODS

This is the graphic presentation of the research procedures (flow) carried out in our research:



2.1. Data collection of GWAS AID association SNPs

The starting point of our research was the collection of publicly available GWAS association data for autoimmune/inflammatory disease phenotypes (AID). All AID GWAS associations were obtained from the NHGRI (National Human Genome Research Institute) GWAS catalog (Hindorff et al., 2009), (http://www.genome.gov/gwastudies). For our study, we selected only AID phenotypes for which the aTNF therapy has been approved: RA, AS, Ps, PsA, CD, IBD and UC.

We used software tool Phenotype-Genotype Integrator (PheGenI), a phenotype-oriented resource for retrieval of data from the NHGRI GWAS catalog, to find all AID SNP associations with the value of $p > 1x10^{7}$. We sorted the associations according to their genomic context as intergenic, intron, nearGene-5 or -3, and coding missense or synonymous SNPs. The selected SNP associations also were labeled with the specific **rs** identifier, their chromosome location, the closest provisional genes linked to them and PubMed identifiers of the first publications indexed for MEDLINE. Identifier “rs” stands for a **r**eference **S**NP ID number, a unique identification tag assigned by NCBI after submission and it does not infer position (NCBI resource).

The associated SNPs reported in AID GWAS are the lead SNPs (often also called tag or signal SNPs). In our research, we analyzed all of them and clearly distinguished whether they are simultaneously functional SNPs. Functional SNPs are defined as SNPs that influence a specific function of a gene or a gene region. Not all lead SNPs are functional SNPs, or vice versa.

2.1.1. Software tools and databases used for data collection

For data collection, we used database NHGRI Catalog and PheGenI software tool.

2.1.1.1. NHGRI Catalog

The NHGRI Catalog of GWAS (Welter et al., 2014) provides a publicly available manually curated collection of the published GWASs. It contains at least 100 000 single-nucleotide polymorphisms (SNPs) All SNP-trait associations found in GWAS cataloged in it have p value equal or lower than 1×10^{-5}. NHGRI Catalog data are extracted from the published literature. Data extraction and curation for the GWAS Catalog is an expert activity; each step is performed by scientists supported by a web-based tracking and data entry system, which allows multiple curators to search, annotate, verify and publish the Catalog data. It is updated regularly on a weekly basis. All GWAS in the catalog conform to eligibility criteria, including number of SNPs assayed and non-targeted study design. The number of SNPs extracted and curated per study is not limited to fifty, as it used to be before 2012. For each GWAS, a trait that best represents the phenotype under investigation is assigned by the curators and it does not rely only on author's submission. We used the NHGRI Catalog version that was available and accessible before the end of November 2014.

Until recently, the traits in the GWAS Catalog were available as an unstructured flat list, allowing only querying through direct string matching. To overcome the Catalog's limiting potential, and establish the Catalog as an integral part of a wider network of genomic resources, several ontologies for mapping the Catalog data have been evaluated; this work is still in progress. Recently, schema ontology was designed to model the key concepts represented in the Catalog, containing all the data in the GWAS Catalog in a format that can be processed by an ontology reasoner and queried. It allows for much richer querying than simple string searching. Mapping to ontologies also greatly facilitates data integration across heterogeneous data sources such as data extracted from the scientific literature. A new addition is the GWAS Diagram, a novel interactive dynamic query interface of traits visualized on the human karyotype, linked to literature and other resources.

2.1.1.2. PheGenI

Phenotype-Genotype Integrator (PheGenI) is the software phenotype-oriented resource for retrieval of data from the NHGRI GWAS catalog (Ramos et al., 2013). PheGenI tool amalgamates content from several NIH resources in addition to the data from the NHGRI GWAS studies: dbGaP, which archives the data of studies investigating associations between genotypes and phenotypes; dbSNP, which includes data on SNPs and their frequencies, position, etc; NCBI Gene, which includes gene-specific data, such as nomenclature, chromosomal localization, gene products; and eQTL data from the Genotype-Tissue Expression Project GTEx.

PheGenI searches items based on MeSH terms (Medical Subject Headings) and retrieves SNP data together with their chromosomal location, gene, SNP ID, and phenotype. Results are provided in a tabulated view and could be downloaded; the results include separate tables with SNPs, genes and associations, a dynamic genomic sequence viewer, and gene expression data if they exist. Data may be sorted according to the source, functional class of SNPs, phenotype, and p values of associations. PheGenI is still under active development and it continues to improve, merging NHGRI genome-wide association study (GWAS) catalog data with several databases housed at the National Center for Biotechnology Information (NCBI), including Gene, dbGaP, OMIM, and dbSNP, all of which we used for our research.

2.2. SNP-gene analyses: mapping AID SNP to genes and other genomic structures

2.2.1. Rationale

It was necessary first to map GWAS AID SNP data to genes or other genomic structures as a step towards understanding of their impact. The simplest and most used method for linking SNPs to genes is mapping the AID SNPs to the nearest genes or genes within the specified distance; however, it is erroneous not to take into account linkage disequilibrium (LD) patterns that might vary with populations and might provide a possibility of finding regulatory sequences (instead of coding genes) within LD regions.

In our study, we rely on LD regions as more relevant and broader determination of the origin of GWAS signals; we did rely only on proximity between SNPs and genes when we analyzed ncSNPs for ncRNAs in deserted regions, as the LD values for those regions are often vague, unknown or too broad.

We prioritize primarily the coding GWAS AID SNPs (cSNPs) to protein genes, because currently it is only possible to incorporate proteins into pathways or networks. However, we also searched for non-protein genes in their respective LD. If no genes existed in high or perfect LD (r^2=1 or r^2> 0.8) with a SNP, we extended search into the regions with lower LD (LD with r^2< 0.8).

Almost half of the retrieved AIDS GWAS SNPs are physically located in noncoding intergenic regions of human AID genomes. The potential effects of these GWAS ncSNPs might not be executed via proteins (although it is a possibility), but rather via regulatory control over biological processes, because alterations in noncoding structural elements of the genome are linked to regulatory control of biological processes. In the case of ncSNPs, especially ones that are located outside introns or completely outside gene regions, in intergenic regions (once called “gene deserts”), we additionally manually searched for annotated structures in the proximity of each ncSNP on the human genome map, upstream and downstream of each SNP up to 200-500 Kb, independent of the size of its LD region.

Unambiguous assignment of disease causality for sequence variants is often impossible, but using the tools and approach as described, we performed SNP-gene prioritization at the current state of knowledge.

2.2.2. Procedure

Using HaploReg v2 for every tag SNP obtained from the NHGRI catalog, we identified all SNPs in strong LD (perfect or high LD: $r^2 > 0.8$), based on 1KG Project data and HapMap European population data, CEU or EUA data. We then searched for all genes in the same high LD region using the same HaploReg tool, correcting if necessary the provisional genes provided in the original GWAS with the newest updated information.

For each AID, we gathered all genes in high LD as potential functional genes influenced by lead SNPs. Genes in LD with any SNP might change with data annotations updates, indicating the importance of the most recent annotations.

In the next steps, all collected genes were evaluated for potential functions using the available data in the Gene NCBI database. Gene roles were detected as described in existing Gene GIFs (gene references into function) data. We used NCBI data primarily, but also Ensembl or Havana project; Havana project often updates annotations outside of the regular releases. Many of the coding genes linked to AID SNPs had no annotations in NCBI databases, nor in other databases. In addition to coding genes, we found that many of the AID SNPs are linked to processed transcripts (mainly various non-coding RNA types) and pseudogenes, some of which might even be processed (translated) (Poliseno et al., 2015). For these structures, we also checked not only NCBI but also other databases such as Ensembl and Havana project (same as for the protein coding genes). Information on ncRNAs are scarce, but even fewer data exist for pseudogenes.

We further examined possible regulatory effects of the AID SNPs (including synonymous cSNPs) with the HaploReg software tool v2 and v4 (Ward and Kellis 2012). Synonymous amino acid substitutions were also evaluated for potential regulatory changes, because, although they do not cause amino acid substitutions, and have neutral effect on protein products, they may influence splicing sites or other regulatory spots within the gene region (Ward and Kellis 2012).

The HaploReg v2 analyses were conducted in two steps. Firstly, we interrogated the lead GWAS SNPs for related high LD sets of SNPs as explained; secondly, we analysed annotations of the expanded SNP sets together with the lead SNP, for their presence within promoters, enhancers, transcription factor binding sites, and expression quantitative trait loci sites. The HaploReg queries data generated by the ENCODE (Encyclopedia of DNA Elements) project (ENCODE 2004; Kellis et al., 2014) for various regulatory elements (e.g., evolutionarily constrained sequences, enhancers, DNAse hypersensitive open chromatin regions, promoters, and 3′- and 5′-untranslated regions).

In addition, we evaluated all sets of SNPs in high LD with the lead GWAS SNPs, by scoring their potential regulatory influence by software tool RegulomeDB (Boyle et al., 2012). We used the 1.1 version of RegulomeDB (Boyle et al., 2012; Xie D. et al., 2013) in order to annotate the AID SNPs with regulatory information.

Scores provided by RegulomeDB for the tested AID SNPs are used to sort all AID SNPs for their relevancy in potential gene regulation. We were able to cover almost all AID SNPs, because currently the RegulomeDB queries are able to identify all common SNPs with the allele frequency > 1% (from 1KG project and/or HapMap3). We selected to further analyse only SNPs with the RegulomeDB scores

of 4 or higher (3, 2, or 1), because they have higher potential to influence regulatory elements. The RegulomeDB scoring is explained in the Supplemental table 2 in the Results section. Scores are ranked based on potential of a regulatory element to modify multiple functions of the region containing the tested SNP.

As a rule, we checked our results over time of one year at least three times for all resulting data, up to the August 2015; we were able to find new information for a few genes (such as IL6R) that we used in our study.

2.2.3. Software tools and databases

2.2.3.1. HaploReg

We used the HaploReg v2/v4 software tool for annotations and prioritization of disease associated risk variants for all SNPs (Ward and Kellis 2012). When a LD threshold is specified (defined with an r^2), results for each variant were shown in a table along with other variants in LD. We tested all noncoding and coding SNPs for their respective LDs with r^2 values ranging from 1 to 0.2, and we reviewed and evaluated all potential genes in their respective LD in order to predict a biological relevancy for each SNP variant in publicly available databases. For LD calculations, the 1KHG Pilot project is used for each population that best matches the ancestry of the subjects.

HaploReg provides annotations from two mammalian conservation algorithms, which are designed to detect constrained sequences across mammalian genomes. In addition, HaploReg v2 is displaying change of the motifs caused by every SNP in the selected LD, thus enhancing further speculation about effects of SNPs on regulatory motifs. The HaploReg also displays if some SNPs are bound by proteins, or are DNAse hypersensitive in many cell types and shows enhancers and promoters impacted by each SNP if data are available.

HaploReg v2/v4 is available on http://www.broadinstitute.org/mammals/haploreg/haploreg.php

2.2.3.2. RegulomeDB

The RegulomeDB is a database that annotates SNPs with known and predicted regulatory elements in the intergenic regions of the *H. sapiens* genome. Currently Regulome DB is querying build 141 of dbSNP and all data at RegulomeDB are mapped to human reference sequence h19. Known and predicted regulatory DNA elements include regions of DNAse hypersensitivity, binding sites of transcription factors, and promoter regions that have been biochemically characterized. Sources of RegulomeDB data include ENCODE project, in addition to NCBI public datasets, and published literature from MEDLINE. RegulomeDB also uses the function information generated by the UCSC Genome Browser (Sayers et al., 2012).

The RegulomeDB server is available at http://www.regulomedb.org, and all ENCODE datasets used in RegulomeDB can be accessed via the ENCODE portal at http://encodeproject.org.

2.2.3.3. dbSNP

The dbSNP database has been established at the National Center for Biotechnology Information (NCBI) to serve as a general catalog of genome variation necessary for association studies with heritable phenotypes, pharmacogenomics, gene mapping, population genetics and evolutionary biology (Sherry et al., 2001). SNPs are the most common genetic variation; they occur roughly every 1200 bp in human chromosomes. The complete contents of dbSNP are available at the public site http://www.ncbi.nlm.nih.gov/snp/.

We used dbSNP 141 build.

2.2.4. Evaluation of missense SNPs impact on proteins coded by missSNP AID genes

2.2.4.1. Rationale

The coding sequence variations (cSNPs) resulting in nonsynonymous amino acid substitutions caused by the AID SNPs (missSNPs) were analyzed by PolyPhen software tool in order to evaluate potential influence of the amino acid substitution on its coding protein conformation and subsequently on its function. Mainly v2 was used, but also v4 with updated annotations, because of our repetitive checking for data and results over time of one year (Adzhubei et al., 2010, 2013).

2.2.4.2. PolyPhen-2 software tool and procedure

We used the PolyPhen-2&4 to evaluate a potential of missSNPs to make a conformation change of respective coding proteins and consequently influence protein functions. The web based software PolyPhen (Polymorphism Phenotyping version 2 and 4) is a software tool, which predicts possible impact of an amino acid substitution on the structure and function of a human protein, employing straightforward physical and comparative considerations.

This was a key evaluation for missSNPs, because it provided us with information about potential change of missSNP harboring protein function that, down the road, might influence pathways and networks. Conformational changes might influence various protein functions such as binding capacity to its interactants (applicable for both pathways and networks) or might induce allosteric modification and impact speed of a pathway, quantity of products and pathway feedback. The detected potential of AID nonsynonymous or missense SNPs to change the coding protein conformation has direct functional consequences on PPI and pathways and networks.

PolyPhen-2 is available on http://genetics.bwh.harvard.edu/pph2/.

2.3. Functional analyses of GWAS AID SNPs: gene-pathway prioritization

The GWAS data enables a shift from individual genetic associations to hypotheses about how the effects of multiple genes contribute to disease susceptibility and expression (Hirschhorn 2009). In order to make this shift, functional characteristics of the AID SNPs data must be understood, and for that reason, they were a major part of our study.

2.3.1. Rationale

Functional roles of retrieved AID SNPs could be analyzed in various ways, but the best and most straightforward approach is to find the interactants of the SNP bearing genes/proteins or genomic structures. Potential interactants then might be organized in directional pathways or less organized structures like networks. Most often, the base of the biological functionality of structures influenced by GWAS SNPs consists of proteins interacting with other proteins.

It is important to stress that both pathway and network analyses have foundation in protein-protein interactions (PPI). Very rarely is any other class of macromolecules, except proteins, included in the prediction of formation of networks or pathways for any given gene.

PPI information can be retrieved from a number of online resources. Fewer resources have their focus on PPI prediction, using a variety of algorithms. However, the integration of both known and predicted (imputed) PPI interactions has been achieved currently only by STRING and ConsensusPathDB software tools; both were used in our study as they have some different characteristics and outcomes.

2.3.2. AID GWAS SNP pathway analysis

2.3.2.1. Rationale

Biological pathway information helps interpret genomics data and gain a more mechanistic understanding of cellular function once we are able to find biological pathways for the SNP gene set of interest.

If an AID SNP risk factor (GWAS SNPs are considered risk factors for a disease phenotype) is related to a certain pathway, then that particular pathway could be considered engaged in the pathogenesis of an AID. Although small number of genes is currently classified into functional conglomerates known to have a certain biological function, once a gene is annotated for a function or assigned to a well-defined pathway, then the function of that pathway become an element of a disease. The pathway should be counted for analyses in the disease phenotype(s), as it obviously might be changed with an AID SNP risk factor it harbors. However, there is no data or models about the flow of pathways currently, and all attributed changes will have to be tested empirically (observation and experiment) or in predictive models. Our knowledge about pathways is singular, and it does not protrude into multifaceted presentation of an pathway.

Biological pathways are collected in databases like KEGG (Goto et al., 1997), Reactome (Joshi-Tope et al., 2005) or the user-curated WikiPathways (Pico et al., 2008) and many other pathway databases. For our analyses, we almost exclusively used KEGG. Occasionally, we used other pathway databases such as NCI (http://pid.nci.nih.gov/) and Reactome (http://www.reactome.org/) as a source for pathway data retrieval or crosscheck.

We relied on the KEGG (The Kyoto Encyclopedia of Genes and Genomes) pathway database (Kanehisa et al., 2012, 2014), because it is considered the largest and most complete pathway database and is regularly updated resource. It is also the most connected pathway database, linked to or used by

many other resources, including NCBI databases, and tools like Cytoscape, STRING, and ConsensusPathDB (all of which we used in our research).

2.3.2.2. Procedure for finding AID SNP pathways

Manually retrieved KEGG pathways and KEGG pathways obtained by pathway enrichment analyses were compared and their potential roles in the pathogenesis of AIDs were evaluated.

2.3.2.2.1. Manual search for AID SNP pathways

We manually searched and retrieved KEGG pathways for all GWAS SNP genes/proteins and TNF. For each AID we compiled the group of KEGG retrieved pathways guided by the disease SNP harboring genes found for each AID. The AID pathways we consider pathogenic or disease pathways for corresponding AID, because they are believed to be engaged in the AID pathogenic process.

2.3.2.2.2. AID SNP pathway enrichment analyses

We also sought to identify disease pathways that are shared amongst AID SNP genes (AID disease genes) by performing pathway enrichment analysis using KEGG pathways database. Pathway enrichment analysis (Mooney et al., 2014) is a technique to find biological, functional sense of any gene dataset.

We used two tools to do the pathway enrichment analysis: STRING and ConsensusPathDB that are explained in the section “Databases and software tools” (below).

2.3.2.2.3. Parsing of the AID SNP pathways

We also wanted to know whether the KEGG pathways, found to home the AID SNP genes, have any common members among them that might be important for the pathogenesis of AID. Finding common members would tell whether KEGG pathways communicate and for that reason influence each other by sharing the same elements. We wanted to find intersections containing subsets of common genes between two disease pathways and explore connections between these common subsets and AID SNP harboring genes. It is intuitive that various pathways could influence each other, especially when they share genes and when they are closely position within the cell loci (membranes, inflammasome, proteasome, etc.).

The only tool that allowed us for this type of analyses was Cytoscape, because both STRING and ConsensusPathDB software do not allow querying for intersections (Su et al., 2014). Cytoscape allows import of major biological pathways from KEGG database (still in development, as it does not contain all KEGG pathways) and then by using the intersecting tool, it finds overlapping sunsets (or common members) between KEGG pathways.

In the process of finding potential AID SNP pathway intersections, we used the CytoKEGG plugin (Bindea et al., 2009) to retrieve known KEGG pathways for TNF and AID SNP GWAS proteins by importing all members of corresponding KEGG pathways. In the second step, we employed tools to intersect two pathways. Thus, we identified common, shared members for each pair of KEGG pathways.

By comparing the intersecting pathway results, we identified proteins common to the analyzed pathways. Their presence indicates the crosstalk among pathways; pathway crosstalk might have significant influence on conditioning of an AID or a response to aTNF therapy.

2.3.3. AID SNP GWAS network analysis

For construction of networks for AID SNP genesets and consequent network analyses, we used Cytoscape software tool. Networks might also be constructed using STRING and ConsensusPathDB tools, but they cannot be dissected, parsed or unionized as with Cytoscape software, nor they can be analysed for mutual members.

2.3.3.1. Cytoscape

2.3.3.1.1. Rationale

Cytoscape is one of the most popular open-source software tools for the visual exploration of biomedical networks composed of biomolecules (proteins, genes) and various types of biomolecular interactions among them (Shannon et al., 2003). Cytoscape is a software suite most powerful when used in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans. It is constantly upgraded as an open source with contributors adding plugins for almost any type of analyses (Saito et al., 2012). Cytoscape provides core functionality to load, visualize, search, filter, and save networks. Many plugins have been contributed by the community to extend Cytoscape functionality and to address specific research needs. The latest generation of Cytoscape (version 3.0 and later) has substantial improvements in function, user interface, and performance (Su et al., 2014).

Cytoscape v3.1.1. is available at http://cytosca-pe.org/download.php, the site hosted at the UCSD (University of California in San Diego).

2.3.3.1.2. Procedure

We initialized networks for TNF and each SNP GWAS gene/protein separately. We imported data and limited these networks only to direct interactants. A typical Cytoscape workflow begins by importing interactions (in our case protein-protein, genetic and regulatory interactions) from public databases for all genes under study. First, we performed keyword searching through 'Import' function for each gene using a gene name or gene ID. Public databases of interactions are accessed using plugins (such as AdvancedNetworkMerge plugin we used that are already standard capabilities incorporated into Cytoscape suite). Plugins allow for online data import, network generation and visualization. During the search step, the databases were selected for data import from protein-protein interaction databases such as Biological General Repository for Interaction Datasets (BioGRID), Human Protein Reference Database (HPRD), Molecular Interaction Database (IntAct), STRING etc. (Szklarczyk et al., 2011, 2015; Kerrien et al., 2012). Then we used Cytoscape to perform network partitioning and comparison between networks by generating Venn diagrams (intersections and unions) We queried the networks for their intersections in order to find common, intersecting shared nodes, actually subnetworks that belong to both networks (Saito et al., 2012). The intersections were analyzed for each network. In order to find

whether any of AID SNP gene nodes interact with each other and in that way discover potential link between AID GWAS genes, we proceeded to construct a union network of the resulting intersecting networks for all available networks. We named this resulting union network the core AID network. The core AID network contained only the GWAS SNP proteins and TNF nodes that are interact among themselves: the nodes linked with the edges indicating a specific type of interaction between each pair of nodes. All other nodes without edges, actually nodes that did not interact among themselves, and did not form a network were eliminated. The edges of all constructed networks signified the interactions between two proteins; they were annotated with details about data source or publications.

2.4. Functional analyses of AID GWAS SNP data using Gene Ontology

2.4.1. Gene Ontology Enrichment Analysis

An alternative way to find functional characteristics of the AID gene sets is by usage of the Gene Ontology (GO) database, which categorize gene using three hierarchical biological categories: molecular functions, biological processes, and cellular components. The enriched GO terms provide information about potential functionality of the tested gene set that is affected in AID.

We employed GO to perform enrichment analyses on the AID SNP gene sets and find the GO terms that are overrepresented for each gene set under study. If enrichment in any of these GO categories exists, it would help us better understand their function. However, although genes assigned to a particular GO category may be associated with similar functions, this grouping does not indicate known relationships or interactions between the genes in each set, and for that reason is less informative that pathway analyses, but more informative than network analyses, that provides interactions.

2.4.2. Rationale

The goal of GO enrichment analysis is to determine which biological processes, molecular function or cellular component terms might be predominantly linked to the affected set of genes in our study (Blake 2013). The simplest approach is to calculate enrichment for each GO term, actually a proportion of genes with certain annotations among the significantly changed genes determined in the AIDS GWAS, when compared to all of the human genes. The set of all genes, or reference set, compasses all known genes in the human genome. An appropriate statistical test must be applied to allow the results to be interpreted as evidence. The analysis process calculates probability as a p-value of the occurrence of a term labeled portion among significant set (sample frequency), versus a term labeled portion among reference set (background frequency). P-value is probability of proportion being annotated with the same term for the tested gene set vs. all genes annotated with the same term. Difficulty in determining significance using the calculated p-value and a cutoff of 0.05 is that multiple testing increases the likelihood of obtaining what appears to be a statistically significant value by chance. Therefore, a correction must be calculated. However, Bonferroni correction for multiple testing is too restrictive and the false discovery rate (FDR) is a better measure for statistical significance in this case. Calculation of FDR, which provides an estimate of the percentage of false positives among the categories considered enriched at a certain p-value cut-off, allows for a more informed choice. Multiple hypothesis testing is a general problem that is not specific to GO enrichment analyses (Farcomeni 2008).

Other limitations of GO enrichment analyses are incomplete annotations (at least 20% of genes do not have GO annotations) and a strong bias towards BP, because GO annotations are not always independent items as conditional for FDR corrections.

2.4.3. Procedure

We used both STRING and ConsensusPathDB software tools to perform GO enrichment analyses.

First, we selected the gene sets for analyses, based on the previous step results: missSNPs, ncSNPs and allSNPs. Each set of genes was queried for GO term enrichment separately, and GO terms for biological process (BP), molecular function (MF) and location in cell compartment (CC) were separately searched, retrieved and analysed.

Both Bonferroni and FDR p values were retrieved and presented. We selected the resulting data mainly based on FDR corrections. In some cases, there were no GO term enrichments after applying Bonferroni correction and FDR; in that case, the results were not presented, because they did not have any significance.

2.5. Software tools and databases used in prioritization and GO enrichment analyses

2.5.1. STRING

The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) (http://string-db.org) (Szklarczyk et al., 2011, 2015; Franceschini et al., 2015) is a software tool and database of known and predicted protein interactions. It has been developed and maintained by EMBL (European Molecular Biology Laboratory), SIB (Swiss Institute for Bioinformatics) and CPR (Center for Protein Research). It is dedicated to provide a critical assessment and integration of protein–protein interactions, including direct (physical) as well as indirect (functional) associations. Interactions between proteins help to describe and narrow down a protein's function; more knowledge is available, less difficult is to find a protein's place in the complex systems. STRING provides the most useful networks that integrate all types of interactions: stable physical associations, transient binding, substrate chaining, information transmission and others.

STRING quantitatively integrates interaction data from many sources for a large number of organisms. It incorporates data from genomic information including data encompassing conservation, similarity, data such as protein domains and protein structures, expression data, concurrence and localization data, high-throughput experimental data and textmining from literature: Medline abstracts and open access articles. STRING holds experimental, but also predicted interaction data. Apart from in-house prediction and transfer algorithms, STRING also relies on many fine resources maintained elsewhere. A list of STRING integrated sources is long and rising. Currently STRING has data on over 5 million proteins and more than 200 million interactions stored.

The current version 10.0 of STRING allows for enrichment detection of human disease associations and pathways annotations, which might be statistically enriched in a given network. In addition, it also allows

enrichment analyses of GO terms enrichment. We used all these features for our analyses: pathway and disease enrichments, and GO term enrichments.

The STRING results can be downloaded or saved as screenshots or tables.

2.5.1.1. Procedure

Starting from the user interface, we have entered the different sets of proteins obtained from the previous steps. Once a protein or set of proteins was identified, we proceeded to the network view, showing nodes as proteins and edges as interaction between them based on evidence. We included all possible evidence available in STRING and we kept adjusted score cutoffs at 0.7 aiming for high confidence results. We did not limit the network size as we wanted to obtain as many interactants and interactions as possible.

Upon switching to advanced mode, we could additionally analyzed enrichment for KEGG pathways and diseases for the given get of AID SNP genes/proteins. The only downside we experienced was limitation of ten or more genes in a set in advanced mode.

We used STRING version 10 for our research. The retrieved data are presented as the screenshots and tables.

2.5.2. ConsensusPathDB

ConsensusPathDB is a meta-database that integrates different types of functional interactions from heterogeneous public interaction/pathway resources data resources (Kamburov et al., 2009, 2011). Through the integration of its Release 30, ConsensusPathDB assembles a comprehensive map of human interactions and pathways.

Although ConsensusPathBD has focused primarily on the integration of existing database resources, its schema might be used for additional manual upload of experimental interactions.

The database contains human functional interactions, including gene regulations, physical (protein–protein and protein-compound) interactions and biochemical (signaling and metabolic) reactions, obtained by integrating such data from source accessible databases including KEGG. Overall, ConsensusPathDB currently contains 41,271 physical entities, 155,432 functional interactions and 2,205 biological pathways in human. It also contains 12,263 unique curated protein complexes (Kamburov et al., 2011, 2013).

The ConsensusPathDB Web interface offers different ways of utilizing these integrated interaction data with tools for visualization, analysis and interpretation. Among other features, the web interface allows for over-representation/enrichment analysis with uploaded identifiers of gene sets. It also contrasts networks based on direct interactions and indirect via second interactants (we used only first interactants). The obtained results may be filtered for cooperation in curated biochemical pathways or co-annotation with Gene Ontology categories. These features we used in our research.

The ConsensusPathDB can be found on http://consensuspathdb.org/. It stays up-to-date with regular content updates and database releases every 3 months.

2.5.2.1. Procedure

Starting from the ConsensusPathDB user interface for overrepresentation analyses of gene set, we entered the different sets of genes obtained from the previous steps. Once a protein or set of proteins was identified, we filtered results for their participating pathways, and GO terms. We mainly preselected KEGG pathways, but we also analysed Regulome and NCI pathways data and other pathways if they did not overlap with the KEGG pathways. For GO terms, we filtered results for four different levels of BP, MF and CC GO terms. We selected only data within acceptable statistical range (FDR less than 0.5).

The advantage of ConsensusPathDB is that it has no limits on number of genes in gene sets and present only FDR statistics. For GO term enrichment analyses, ConsensusPathDb allows for filtering parent-child levels of complexity. It also shows protein complexes in addition to pathways and networks for a test gene(s) and occasionally introduces gene-gene and gene-protein, as well as drug-protein interactions as regulatory links. Because this option is still in development, we checked it often, but did not present results in this study.

2.5.3. DiseaseConnectDB

The AID pathways were also retrieved and analysed by DiseaseConnectDB software tool. The obtained results were crosschecked with the results of pathways we found based on GWAS SNP data by other methods (previously described). It was used as another tool with different approach that has confirmed our results.

DiseaseConnectDB is a public web database and server based software tool that focuses on the analysis of common molecular mechanisms shared by diseases by integrating comprehensive omics and literature data (Liu et al., 2014). DiseaseConnectDB is currently one of a very few databases that touches on the concepts of disease connections (Rappaport et al., 2013; Peng et al., 2013*).* It is the most comprehensive resource documenting the shared molecular bases of diseases. The DiseaseConnectDB web server contains 18,707 disease–disease, 660,985 disease–gene, 12,617 drug–gene and 113,498 drug–disease relations; these data cover 4,791 diseases, 6,215 drugs and 15,182 genes. It has incorporated a large amount of GWAS catalog, gene expression data, microRNA expression data and text-mined knowledge to discover disease–disease connectivity based on molecular mechanisms.

We used DiseaseConnectDB software tool to crosscheck our results on the AID SNP pathways.

2.6. Finding interactions of AID GWAS SNPs with non-coding RNAs

2.6.1. Rationale

Functional RNA molecules have recently emerged as important regulators of gene expression. MicroRNAs (miRNAs), together with small interfering RNAs (siRNAs), long noncoding RNAs and other categories of non-coding RNAs, are all in this category that has been intensively researched in the human genome. MicroRNAs (miRNAs) are small evolutionarily conserved regulatory RNAs that modulate mRNA and regulate gene expression at the posttranscriptional or translational level. There is evidence that GWAS SNPs may influence miRNAs structure and function (refs).

Therefore, we sought to find whether any of the AID GWAS SNPs are able to change sequence of miRNAs or interfere with the binding site of any regulatory miRNA in gene regions. If an AID SNP occurs in a miRNA gene, then it could affect its regulatory function and help explain association a SNP role with AID. Additionally, if an AID SNP occurs in the docking region of an AID SNP gene and change the docking sequence for miRNA, that SNP might have impact on regulatory function of a miRNA. Both options were explored using relevant miRNA databases and tools.

2.6.2. Procedure

Non-coding RNA (ncRNA) genes were manually searched for around GWAS AID SNPs, up to 200 bp upstream and downstream of each SNP locus starting from the dbSNP, and in the corresponding SNP high LD regions as determined by HaploReg v2 software tool.

For this experiment, we selected the top hits among the GWAS AID SNPs, ones that might be considered to have the highest probability of influencing ncRNAs. That includes the SNPs with the highest p values in the GWAS and the SNPs with the highest scores obtained by RegulomeDB.

In order to find out whether any of the selected AID SNPs interfere with miRNAs, we sought to find the AID SNPs located in miRNA genes. We also examined whether any AID SNPs interferes with known target sequences of miRNAs, where a SNP can change the docking sequence for miRNA in a way that influence its binding.

We searched the microRNA database miRBase for the miRNAs found in high LD with the preselected SNPs. The miRBase (Agarwal et al., 2015) database is a searchable database of published miRNA sequences and annotation and is found at http://www.mirbase.org.

We have also searched miRNA targeting sequences in the regions of the AID SNP genes. We examined all AID SNP gene regions for evidence of co-location of SNPs within its boundaries of AID SNP genes using TargetScan tools. We employed the TargetScan Release 7.0 that can be found at (http://www.targetscan.org/).

2.7. Overview of the bioinformatics software tools and databases used in the study

Tool / Database	Relevance	Version
GWAS NHGRI Catalog	Databases with all GWAS studies; annotated by professional staff; publicly available	2013-ongoing
PheGenI	Search and retrieve tool for queried disease SNPs from association studies in the NHGRI Catalog	still in construction
HaploReg v2	Detection of SNPs in LD with a SNP of interest; detection of genes in LD with a lead GWAS SNP	Versions 2 and 4
RegulomeDB	Scoring for potential regulatory function of the relevant GWAS AIDs SNPs	Version 1.1
dbSNP	SNP database; SNP characteristics, location; MAF	Build 141
KEGG database	Detection of genes of interest; characteristics; known relations to diseases and their pathways; classification of pathways and relation to modules	Continuous release
Entrez Gene NCBI	Search for genes and their properties: gene function and location in the chromosome or region	Build 141
Cytoscape: Network construction and analyzer AdvancedNetworkMerge plugin	Construct networks for genes that are related to the GWAS AIDs SNPs; analyze the same networks for unions or intersections	v3.1.1
Cytoscape: KEGG pathway query tool and analyzer CytoKEGG plugin	Analyse relationship between KEGG pathways	v3.1.1

STRING	Gene prioritization; pathway enrichment analysis	Version 10
ConsensusPathDB	Gene prioritization; pathway enrichment and complex formation analysis	Release 30
GO ontology	Retrieval of GO terms for gene enrichment (BP, MF, CC and pathways); potential function assignments	Version 4.0
DiseaseConnectDB	Retrieval of pathways common to AIDs	Version 1.1
miRBase	Searching for published miRNA sequences and annotation	Release 21
TargetScan	Predicting targets for microRNAs	Version 7.0
DrugBank	Characteristics of TNF biologics; interactants, inhibitors, targets	Version 4.3
PharmaGKB	Characteristics of TNF biologics	Version 4.0

3. RESULTS

3.1. AID-associated single nucleotide polymorphisms (AID SNPs)

We retrieved 383 SNP-AID associations from the NHGRI GWAS Catalog collection by the end of November 2014 (Table 1, Supplemental Table 1 and Supplemental Figure 1). These AID SNPs were associated with 356 unique genes. Some AID associated SNPs are shared by two or more genes, thus resulting in a many to one SNP-gene association. The AID SNPs were dispersed throughout all somatic chromosomes; sex chromosomes (x and y) did not contain any AID SNPs (Ideogram, Supplemental Figure 2.).

Genomic context for each retrieved SNP is provided in the Supplemental Table 1; the genomic context distribution of the SNP associations is provided in the Table 1, alongside with the numbers of affected genes for each AID, and in the chart (Supplemental Figure 1.).

For each association the SNP identification number (rs), location on a chromosome and a source of information are given in the Supplemental Table 1. More than half of all AID SNPs are intergenic, 30% are intronic and only 7,5% are missense SNPs; 3% are UTR-3 and UTR-5 SNPs, 5% are nearGene-5 SNPs and 1% are nearGene-3 SNPs; coding synonymous SNPs represent only 2,5% and there are only 2 frameshift SNP among retrieved GWAS associations.

Table 1. AID-associated SNPs and genes for each disease

	Number of					
Autoimmune/inflammatory diseases (AID)	**SNPs**	**Genes**	**miss SNPs**	**nearGene SNPs**	**intronic SNPs**	**intergenic SNPs**
Rheumatoid Arthritis (RA)	102	73	4	6	34	58
Arthritis, Psoriatic (PsA)	5	8	1	1	0	3
Psoriasis (PS)	35	39	5	1	14	15
Crohn Disease (CD)	96	120	7	5	37	47
Inflammatory Bowel Diseases (IBD)	21	15	1	0	13	7

Spondylitis, Ankylosing (AS)	26	35	2	1	8	15
Ulcerative Colitis (UC)	98	52	7	9	21	61

AID associated GWAS SNPs: number of SNP for each AID organized according to their context and number of genes influenced by the same SNPs as provided in the NHGRI Catalog on November 2015.

3.2. Gene candidate identification

We performed SNP-gene prioritization experiments to assign both coding and non-coding SNPs to their most probably associated genes.

We first identified and verified all AID SNP variants using the dbSNP database. All AID SNPs have been recognized in the 1000 Genome project and HapMap3. We manually verified the genes assigned to the AID SNPs using the dbSNP database and HaploReg v2 software. We found that several genes assigned originally to the AID SNPs were not exactly located in the SNP regions according to the newest information from HapMap3 and 1000 G project. We used further the updated, corrected version of genes associated to the AID SNPs.

The AID SNPs from HLA regions were excluded due to the intrinsic high variability of HLA regions crammed on the chromosome 6 p 31.21. section and in perfect LDs (r^2=1) over the long stretches of chromosome in the same region.

3.2.1. Identification of gene candidate dataset for coding non-synonymous AID SNPs (missense SNP gene dataset)

We identified 23 missense SNPs among significant ($p < 1 \times 10^{-7}$) retrieved AID SNPs (Table 2). All missense AID SNPs have been detected in the 1000 Genome Project and HapMap3. The minor allele frequency count (MAF) is given for each SNP. It shows all missense AID SNPs are common in the human population, at least in the population of European origin (usually labeled as CEU or EUR), for which the majority of AID GWASs had been performed.

We find that only 3 out of 23 missense SNPs were present as the solely risk variants associated with a particular AID. The majority of missSNPs were risk variants for multiple AIDs under study. The same missSNPs were also the risk factors for other immune/inflammatory diseases such as SLE or T1D (as found in the NHGRI catalog database) (Table 2).

This finding suggests the pleiotropic relations among missSNPs, their genes and AIDs.

3.2.1.1. Missense AID SNP dataset evaluation

The characteristics of GWAS AID missSNP are provided in the Table 2 and 3. Each missense AID SNP (missSNP) is labeled with rs number, minor allele frequency count (MAF), presence in the 1000 Genome Project, and the type of amino acid alteration caused by a nucleotide switch. Each missSNP the associated gene is given labeled with corresponding official symbol and NCBI gene identifier. Also presented is the AID in which the missSNP is found alongside with other autoimmune diseases (other than AID). Relationship of the missSNPs with other SNPs in the same gene is commented (if they are known in the NCBI databases).

Table 2. Missense GWAS AID SNPs characteristics: part I
(amino acid changes caused by missSNPs, PolyPhen-2 evaluation of functional consequences)

We sought to predict the functional impact of the AID missense SNPs on proteins coded by the genes harboring these missSNPs. We used the PolyPhen-2 to predict their functional impact, and the results of the experiments are presented in Table 2. Results of amino acid change evaluation are given as "damaging" if the amino acid change is highly probable to alter protein conformation to the point of damaging its function, or "benign" if the conformational alteration is not damaging for protein function based on current knowledge. Roughly at minimum, one third of the missense SNPs (8 out of 23) were evaluated as damaging for the function of corresponding proteins coded by the genes harboring the missSNPs.

An illustration of the obtained results for evaluation is provided in the Figure 1, where for instance, missSNP rs8192591 is predicted to be damaging (Figure 1). Each missSNP evaluation has been repeated at least several times.

Figure 1. An example of functional evaluation of missense SNP impact by PolyPhen-2

To assess the potential of the missense SNPs to modify gene regulation and expression of the genes in the affected regions, we analyzed the high LD haploblocks for each missSNP using RegulomeDB before August 2015 (Table 3).

Six SNPs (out of 23 missense SNPs) with the highest scores (1a-3b) were found. The scores suggest that rs2476601, rs4077515, rs3764147, rs3184504, rs20541, and rs2305480 may significantly modify regulatory function of their corresponding DNA segments in high LD. The rest of the missSNPs were

scored 4-6 (or no data existed at the time these analyses were carried out), indicating that they were less likely or not at all able to influence regulatory functions.

The majority of missSNPs (14 out of 23) belongs to the regions that are highly conserved across the mammals, even conserved in the vertebrates (Table 3). The majority of missense SNPs variants are able to modify several regulatory DNA elements including regions of DNAase hypersensitivity, promoter and enhancer histone marks, transcription factors binding sites and promoter regions.

Among the observed motif changes, there is an affected NF-kB motif in the region for ERAP1 gene, while the TRAF3IP2 gene region contains a STAT motif changed due to the presence of rs33980500 variant. Only 2 missSNPs were not evaluated as expression quantitative trait loci (eQTL), while the majority had multiple eQTL results. Almost each missSNP has several regulatory markers (Table 3).

Table 3. Missense GWAS AID SNPs characteristics: II part
(analyses of LD blocks by HaploReg v2 and RegulomeDB scoring)

Further, we analysed LD regions around the missSNPs for other SNPs and genes. Potentially, other genes might also be reflected or represented by the tagged missSNPs on used GWAS chip platforms. Alternatively, other SNPs in high LD with the tagged missSNPs might also be functional SNPs and acting in accordance with the tagged missSNPs, but were not detected on the used GWAS platforms because the used chip platforms did not recognized them.

We found that nearly half of all missSNPs (11 out of 23 missSNPs) were in high LD with additional SNPs in the same gene, and at least nine missSNPs were in high LD with SNPs belonging to other genes (Table 3). This suggests that high or perfect LD missSNP blocks show existence of additional SNPs and alternative genes for majority of AID missSNPs.

We were able to distinguish between lead (tagged) and functional SNPs among the set of 23 missense SNPs based on orthogonal evidence. The missSNPs belonging to MST1, MICA, YDCJ and GSMBD genes were not functional SNPs. Several other missense SNPs, all assessed as functional, had additional functional SNPs in very high LD, including SNPs within the boundaries of another gene. This was case for CARD9, SNAPC4, NFKBIE, EGFL8, GPSM3, IL6R, IL17REL, and IL7R genes; these SNPs might have additional functional consequences.

We found that some lead missSNPs were in high LD with another missense SNP in the same gene. These other SNPs were not recognized by the used GWAS platforms (Table 3). However, their existence might influence and change functional characteristics of missSNPs and genes they harbor, but

it is not possible to evaluate them together, because there is no adequate tool for collective evaluation of SNPs at the same time.

For example, the NFKBIE gene has a missense SNP rs2233434 detected as a lead, in perfect LD (r^2 =1) with another NFKBIE missense SNP rs2233433. Their independent damaging potential on protein sequence and regulatory function potential is not high (scored 4) when calculated for each variant separately based on the PolyPhen-2 and RegulomeDB tools. However, they might be more significant variants when both are present simultaneously (both are common SNPs, both are missense SNPs and both are in perfect LD, meaning inherited simultaneously).

Similarly, GSDMB was found to have two missense SNPs also in high LD. One missense SNP (rs2305480 discovered by the AID GWAS) was potentially damaging for GSDMB protein's conformation and function, and the other was not; however, both GSDMB missSNPs have high scored potential for gene expression modification (1f score), meaning they both may have a strong effect when present.

There is no software tool currently that evaluate simultaneously two or more SNPs, even missSNPs that change amino acid sequence, nor to evaluate changed DNA sequence as for synonymous coding SNPs, because it would require structure-function knowledge that does not exist yet.

3.2.1.2. Assessment of additional genes in LD with missense SNPs

Nine missense SNPs (out of 23 missSNPs) had alternative functional genes in high LD (r^2>0.8). We assessed these additional genes for potential roles in AIDs, because the lead missSNPs might reflect their influence on AID as well.

We examined these additional genes in the context of their function within the immune system, presence in the immune system pathways and association with other autoimmune/inflammatory diseases. The gene evaluation and selection process based on these criteria and its results are presented in the Table 4.

Table 4. Alternative coding genes at the missense AID GWAS SNP loci

Only the alternative gene UBE2L3 appears to have sufficient support for an AID-related role. UBE2L3 gene encodes a member of the E2 ubiquitin-conjugating enzyme family. This enzyme is demonstrated to participate in the ubiquitination of p53, c-Fos, and the NF-kB precursor p105 *in vitro*. The modification of proteins with ubiquitin is an important cellular mechanism for targeting abnormal or short-lived proteins for degradation.

For all other high LD genes, based on our criteria we cannot argue that the alternative genes are substantial genes. However, because the data for the analysed genes is so scarce in all current gene databases, we could not select more genes with certainty. However, it is obvious from the results provided in the Table 4 that even the selected genes are selected based on partially convincing criteria, because the alternative is simply even less known (example is NOTCH gene vs EGFL8 gene, or ATXN2 gene etc.). The selection is made solely based on currently available data, which may change.

Nearly all AID missSNP exhibited at least one of the characteristics with profound functional consequences, so they might be considered functional on several levels: influencing coding genes, influencing other genes in LD, or disturbing regulatory function in their regions.

3.2.2. Identification of gene candidate dataset based on non-coding AID GWAS SNPs (ncSNP gene dataset)

The majority of AID GWAS SNPs appear to be located within intronic or intergenic regions (Supplemental Table 1 and Supplemental Figure 1). This makes it challenging to assign functional roles to these variants and explain their disease causation. Based on available data, we examined non-coding AID SNPs functional relevance by searching for genes (coding or non-coding) and regulatory elements associated with them.

3.2.2.1. Evaluation of non-coding SNP dataset

The non-coding AID SNPs with the p value smaller than $p<1 \times 10^{-7}$ value were retrieved from the AID GWASs (Supplemental Table 2).

To help separate potential functional non-coding variants from the large pool of the AID ncSNPs, we used Regulome DB and HaploReg-v2, along with data from the HaploReg v2, NCBI dbSNP and the EntrezGene NCBI databases.

We thoroughly searched the ncSNP LD blocks identified by HaploReg v2 for genes, checking the regions not only in a perfect LD, but also regions with medium stringency of linkage disequilibrium (0.8-0.6 $<r^2>0.4$).

To test for regulatory functional potential of ncSNPs, we assessed chromatin status of the ncSNP LD blocks for potential for regulatory function by the Regulome DB. The RegulomeDB scoring data for all ncSNPs are presented in the Supplemental Table 2:

Supplemental Table 2. RegulomeDB scoring results for GWAS AID SNPs

Almost 25% of all AID ncSNPs obtained the highest scores (1f-3b). This indicates their significant capacity to influence gene regulation. The ncSNPs that scored 4-6 were evaluated as less likely to modify gene regulation and were not further examined.

The high-scored SNPs were further analyzed and their characteristics are presented in the Table 5, classified by the associated AID:

Table 5. Characteristics of highest Regulome DB scored GWAS AID ncSNPs

Almost all high-scored ncSNPs were associated with at least one coding gene, non-coding RNA genes (ncRNAs) or pseudogenes. Only six out of more than fifty highly scored ncSNPs were located in the regions with no known genes. Consequently, the majority of highly scored ncSNP have some kind of potential functional influence in the AID's genes.

The genes in LD presented in the Table 5 somewhat differ from the genes reported in the original GWAS studies for the same ncSNPs. We believe that our selection is more accurate, because we identified these genes with more precision, using updated, richer and more precise annotations from the Human 1000 Genome Project data and HapMap3 data, than ones existed at the time of a GWAS publication.

For example, a risk variant for IBD is the high scoring ncSNP rs8049439, and it is associated with ATXN2L (a gene that encodes a protein of unknown function) and with TUFM (a gene that encodes a protein that participates in protein translation in mitochondria). Neither has any known links to immune system functions. In addition, this ncSNP is in high LD with an microRNA, also with no data about its relevancy in the immune system.

The majority of the dozen highly scored ncSNPs UC risk variants are linked to immune system functions. One of them, rs3024505, linked to IL10 gene, and is also a risk variant linked to multiple autoimmune/ inflammatory diseases.

We have found that ncSNP rs9263739, scored 1f, was in a perfect LD with the two missense SNPs in the region of the Psoriasis susceptibility 1 candidate 1, 2 and 3 (PSORS1 C1, C2 and C3) genes: two protein-coding genes, C1 and C2, and an ncRNA gene, C3. This SNP was originally detected as a very high risk SNP for UC (p value = 4.000×10^{-67}), located in the region within the gene CCHCR1. Our results negate involvement of CCHCR1 gene. According to our analyses, rs9263739 signals PSORS1 candidate genes C1 and C2, actually their two missSNPs, one in each gene region. Both missSNPs are evaluated by PolyPhen-2 to be benign (results not shown). PSORS1 candidate genes are also found to

be associated with psoriasis in the GWASs containing risk intronic SNPs for C1 and C2 (Table 1). They are recognized by OMIM as risk genes for psoriasis, but not for UC. We have no explanation why CCHCR1 gene is still recognized as a UC linked risk gene instead of PSORS1 C1 and C2 genes, except that they are all situated on chromosome 6, a chromosome notoriously difficult to resolve because of many shorter haploblocks and strong LDs along it.

The same rs9263739 SNP is as an example of a lead vs a functional SNP: it is a lead SNP, but it is not a functional SNP, because it is in perfect LD with two other missense SNPs of PSORS1C1 gene; missense SNPs are by definition functional SNPs. Interesting enough, none of these two missSNPs was among the detected GWAS AID missSNPs. Both exist in 1000G Project.

Another example of clarification of non-coding SNP location was the rs9858542 variant, assigned to gene BSN. However, this SNP is in high LD with already known missense SNP rs3197999, which belongs to MST1 gene. For that reason, it is a lead but not a functional SNP for this region; we already analyzed the functional SNP missSNP rs3197999 of MST1 gene and found it not functionally damaging for the MST1 protein (Table 2).

For rs1297265, we were not able to find any known genes in LD, so we were not able to confirm its links with NRIP1 or CYCSP42 genes. The same was true for rs3806308, similarly with no genes in LD. Our SNP-gene prioritization analyses of ten highly scored RA ncSNPs showed interesting results. Several high scoring ncSNPs have links to genes exclusively associated with immune system functions, like CD40, IRF5, RBPJ, CCR6 and CTLA4, but others have links to genes with no specific role within the immune system, like TNPO3, encoding a tRNA import factor, APOM encoding apolipoprotein M, PDE2A a phosphodiesterase or NCOA5, nuclear receptor coactivator. The latter have roles that are more general in the biosystem.

For example, rs805297 is associated with CSNK2B: this gene encodes a casein kinase, a ubiquitous protein kinase, which regulates metabolic pathways, signal transduction, transcription, translation, and replication, while its beta subunit serves regulatory functions. CSNK2B has no exclusive role in the immune system as it participates in ribosome biogenesis in eukaryotes, tight and adherens junction formation. However, as such it is involved in the immune system signaling: Toll-like receptor signaling pathway, BCR signaling pathway and NF-kappaB signaling pathway, and also in infection diseases (as per KEGG and NCBI Gene databases).

We found that other genes associated with ncSNPs have no known function, although their association with AIDs might be explained with their physical proximity to immune genes or location. For examples,

gene PRRC2A has no known function, but it is localized in the vicinity of the genes for TNF alpha and TNF beta. Gene LY6 encodes lymphocyte antigen 6 complex, also with no known function, but experimentally detected in lymphocytes.

Among 12 high-scored ncSNPs associated with CD, majority is linked to the genes with no immune functions. SMAD3 gene encodes one of the SMAD proteins that act as signal transducers and transcriptional modulators and mediate multiple signaling; including signaling in cancer; however, it does not have any specific function in the immune system. HORMAD2 also has no known function. CCDC88B encoded coil-coil containing protein may be involved in linking organelles to microtubules. Genes SLC22A4 and SLC22A4 belong to solute carrier family 22 of organic cation transporters, which are critical for elimination of many endogenous small organic cations. None has a specific links to the immune system.

Psoriasis was linked to 15 high-scored ncSNPs, with the striking incidence that majority of them are located on the chromosome 6 (C2, SKIV2L, NOTCH4, GPSM3, WASF5, RNF5, DDR1, PBX2, TCF19). Many of these highly scored ncSNP genes have roles in the immune system. Almost half of ncSNPs belong to the HLA complex genes. Other associated genes are predominantly located very close to HLA complex, like genes DDR1 and MUC22, but have no assigned function.

Gene C2 functions in the immune system as one of key component of humoral immunity. It is connected with some autoimmune phenomena.

SKIV2L encodes an endogenous antiviral helicase that participates in RNA degradation and acts in an early phase of immune response.

WASF5P is pseudogene and belongs to the family of genes encoding Wiskott-Aldrich syndrome (WAS) proteins. Wiskott-Aldrich syndrome is a very well recognized disease of the immune system, an immune deficiency. WASF5P pseudogene, which apparently might not be transcribed nor translated, resembles WASF3 protein that is involved in the transmission of signals to the actin cytoskeleton and in the Fc gamma receptor mediated phagocytosis.

PSORS1C1, 2 and 3 genes are dominantly present among highly scored ncSNPs associated with psoriasis, as three loci out of 15 analyzed ncSNPs are linked with the complex. The PSORS1C complex is also connected with several diseases including AIDs; PsA is also linked to the PSORS1C complex, but so far, the p values of associated SNPs were higher than our chosen threshold and for that reason not included in the study (data not shown).

The ncSNP gene dataset, obtained after ncSNP-gene prioritization analyses, consisted of several completely new genes, but also genes previously determined in the missSNPs dataset (such as NOTCH, MST1 and UBE2L3). Prior to determining the final ncSNP dataset, we eliminated the HLA genes due to their complexity and extremely high LD (same as for the missSNPs dataset) and already detected missense SNP harboring genes. We also did not take into consideration pseudogenes, as they are not usually translated genes and have no information regarding potential function in the immune system (or any other process).

We used the final set of ncSNP gene in further functional analyses, in parallel with missSNP gene dataset.

3.2.2.2. Functional assessment of non-coding RNA genes

We extended our search for non-coding RNA genes in the neighborhoods of the top AID GWAS SNPs. We selected sixty top AID GWAS SNPs, ten for each AID, solely based on their p values as the most influential and relevant risk SNPs. The regions around these SNPs were analyzed for ncRNA genes in a similar fashion as previously done. We used HaploReg v2 to determine high or perfect LD blocks, Regulome DB to score their regulatory function potential and NCBI dbSNP to find annotations and provide map enabling manual search for non-coding genes. Results of the analyses are presented in the Table 6:

Table 6. Non-coding RNA genes in LD with top intergenic GWAS AID SNPs

Most ncSNPs have RNA genes in high LD, even several different types of ncRNA genes, including micro RNAs (miRNAs) and pseudogenes.

A subset of 8 SNPs with the lowest p values (8 out of 53 analyzed), were also high RegulomeDB scoring SNPs, thus indicating their very high potential to significantly change gene expression of the targeted genes, or to influence regulatory functions of the region (data collected on Feb 3rd 2015.).

Amongst these eight high-scored ncSNPs, only two loci did not have any ncRNA genes in LD. More than a half of loci contained miRNA genes, and often had two or more different kinds of ncRNA genes in the same region, in addition to miRNAs.

3.2.2.3. Functional assessment of microRNA genes

A set of twelve miRNAs genes in LD with the top ncSNPs were further examined to determine their functional relationship with AID. The results of miRNAs analyses were presented in the Table 9, divided in three sections: a, b and c.

Table 7. Characteristics of miRNA in high LD with AID ncSNP dataset

Section 7a

The majority of the miRNAs were already known human miRNAs with their specific targets (Table 7a). However, none of miRNAs targeted the AID genes.

Section 7b

We also retrieved all miRNA that target 3'-UTRs of the AID genes (Table 7b). None of the twelve miRNAs in LD with the top AID ncSNPs were overlapping with the miRNA detected in the UTR-3' regions of the genes harboring missense AID SNPs.

Section 7c

A few ncSNPs were located in the 3'-UTR- loci of the targeted genes (Table 7c). To check whether we missed any other miRNA that might be altered by AID SNPs, we retrieved all SNPs that belong to 3'-UTR regions, irrespectively of genes they belong. Then we inspected these regions for miRNAs. Repeated comparisons between two sources did not reveal any overlapping data among miRNA targeting AID SNP genes and miRNA linked to AID ncSNPs.

3.3. Functional analyses of GWAS AID SNP gene set: gene-pathway prioritization

All functional analyses were performed based on the knowledge that majority of genomic components or genes must be incorporated into some type of functional units or biological processes.

After the SNP-gene prioritization analyses, we were able to define the final datasets of the AID SNP harboring and SNP influenced genes (Table 8).

Table 8. AID gene datasets: missSNP genes and ncSNP genes

The SNP datasets contained only coding genes, as the existing tools for functional analyses can only connect coding genes with pathways and networks. We kept the gene sets separately as two sets,

missSNP gene set and ncSNP gene set, because missSNP gene set was determined with much more certainty than the ncSNP gene set. We did not want to test all genes together as pooling might undermine efficacy of finding the various enrichments (according to our results shown below), although there is no reason to make one set.

We performed the following functional analyses of the GWAS candidate AID SNPs: gene-network prioritization and gene-pathway analyses, GO term enrichment analyses (for BP, MF and CC) and disease-pathway enrichment analyses. Except for GO enrichments, all used software tools were based on known or imputed protein-protein interactions (PPI). We used several bioinformatics tools for the functional analyses: Cytoscape, ConsensusPathDB and STRING.

3.3.1. Network analyses

3.3.1.1. Network analyses using Cytoscape software tools

The networks for each missense SNP harboring gene and TNF were constructed using Cytoscape (Supplemental Figure 3.). The structured SNP networks consisted only of primary (or first) known or predicted interactants for each protein of interest. The protein-protein interaction (PPI) networks of the missense SNP genes and TNF PPI networks are shown in Figure 2. There were no data available (at the time) to construct the networks for several proteins encoded by GSDMB, IL17REL and IL7 missense SNP harboring genes, and for that reason they had to be omitted from the analyses.

Supplemental Figure 3. Network images for missSNP harboring genes/proteins and TNF

The constructed networks consisted of known or predicted interactants (gene/protein) indicated by the nodes: a cloud of nodes around a central node representing SNP harboring gene/protein or TNF; the nodes were connected by edges representing detected biological interactions between two nodes based on predictive methods (as provided in Methods). The edges signify biological interactions independently of a method by which an interaction was established: experimental biochemical results, text mining data, expression patterns, co-expression and co-localization data, and imputed interactions based on similarity of genes or proteins. In addition to human gene/protein networks, the corresponding networks in mouse, rat (red and white networks in the Supplemental Figure 3.), and occasionally in other organisms, if data were available and in case the orthologue annotations of the same proteins were known in these species and interactions between them were identified.

The biggest network was created for TNF gene: it included over 1500 nodes representing interacting proteins (primary interactants) connected with edges (Figure 2a). The networks constructed for the

missSNP gene/protein contained much smaller number of the nodes; they varied from several nodes to several dozen of nodes. All nodes were connected with edges representing biological interactions between pairs of nodes, a central node and first interactants' nodes. An example of missSNP harboring gene such as NFKBIE gene/protein is presented in the Figure 2b.

Figure 2a: TNF network

Figure 2b: NFKBIE network

We queried the missSNP gene networks to find common, intersecting set of interacting proteins (overlapping members between two networks) among the GWAS AID SNP harboring protein networks, as well as between each GWAS AID SNP harboring gene/protein network and the TNF network. The number of common interactants in the intersections between the constructed networks varied greatly, from no interactants to a very few interactants, up to a dozen or more of common interactants in some intersections of the paired networks (empty intersections are not presented).

Figure 3a. Image of intersection between missSNP ERAP1 and TNF networks

Figure 3b. Image of intersection between missSNP NFKBIE and TNF networks

The resulting intersections were not structured in the form of networks; instead, they were presented as collections of unconnected nodes. There were no edges that would indicate any type of biological interactions among these nodes, except for a very few edges detected only in several instances. In case that common intersecting set of genes existed in other species, they were also collected and presented. Next, we examined interactions among SNP harboring gene/protein networks by taking the union of networks. The resulting network is shown in Figure 4.

Figure 4. Union of GWAS AID SNP harboring protein networks

The union visualization shows how the missSNP genes in AIDs have a potential to interact with each other via direct known or unknown, imputed neighbors (or indirect) intermediators. In the union only the first interactants were preselected and presented; yet the union is a very complex structure, but still a single structure. There are no separate networks, as all networks are interconnected. The union network is suggestive of a space for interactive communications among the SNP harboring proteins that might be important in AIDs pathology.

The number of intersecting nodes representing common members among SNP networks and between SNP networks and TNF network is shown in the Table 9 for easier view and comprehension.

Table 9. Intersections between missSNP harboring genes/ proteins networks and TNF networks

These results indicate how connected the paired genes/proteins are: the greater the number of interactants, the better connected the two genes may be.

As the Table 9 shows, the intersections of the networks belonging to the missSNP genes/proteins and TNF network have a significant number of common interactants. The very existence of the common members, especially in such high numbers, could not have been detected without this computational approach (they are all first interactants filtered by the Cytoscape tool). Here, we show for the first time that majority of the SNP gene/protein networks do have a number of common intersecting members. Especially rich intersection datasets were obtained for the TNF network intersecting with TYK2, TNFAIP3 and NOD2 gene/protein networks. TNF and IL7R also have several common members, including TNF (not shown in the Table 9). The common intersecting members show how relatively central the TNF gene is in relation to the AID SNP harboring genes, because it is almost always present in every intersection, or at least, one of its receptors (TNF receptors) is present.

In addition, there were very rich intersections containing couple of dozens of genes between TYK2, PTPN22, IL23R, CARD9 and TNFAIP3 networks. Since the majority of these genes/proteins are functionally damaged by the very same AID missSNPs (called risk variants for AIDs), their networks might have a ripple effect on all other networks because of their changed protein functions. Only RTKN2 and LACC1 SNP harboring genes/protein networks did not have any common interactants with TNF or other missSNP networks. This finding could have reflected simply a lack of knowledge about their roles and their poor annotation in the databases (which are evident), and not necessarily the lack of potential interactants. However, they might also serve as a negative control for the Cytoscape process.

It was noticeable that some intersections contained the same gene subsets repeatedly. The repetitive common members in the intersection subsets indicated subnetworks, with not yet well defined common functionalities, like unknown pathways, or parts of unknown pathways (because if they were defined, they would appear as nodes with edges, or networks).

We further explored a possibility that all intersecting sets, obtained from querying missSNP harboring gene/protein networks and TNF network, had some subnetwork in common. We searched for the common interactants among all missSNP networks and TNF network by a process of unification of their intersections. When each SNP harboring gene/protein network was queried for unifying set of genes by the Cytoscape tools, the following result was obtained (Figure 5):

Figure 5. Network union between intersection datasets of missSNPs and TNF networks

As a result of unification of the missSNP genes/proteins and TNF networks intersections, we found a very simple network of several nodes with edges. This small union network consisted of the nodes connected with edges representing only TNF and eight missSNP harboring genes: CARD9, IL23R, SH2B3, NOD2, TNFAIP3, TYK2, NFKBIE and MICA. The connecting edges in this union network were based on predictive text mining developed by capturing knowledge from the published literature. These results indicate that, although the direct interactions between missSNP harboring genes/proteins and TNF have not been recognized, and interconnecting paths were not (yet) discovered, the full network could be constructed on imputed data. While the edges have not exposed any experimental data (as they were imputed results based on text mining), nor any particular publication confirming interconnection between GWAS missSNP genes/proteins and TNF was disclosed, the resulting network made a rational base for elucidating the AID GWAS SNP roles. This union network implied a potential for anti-TNF action via these edges, and suggested the factual base for a search of the key gene/protein players responsible for anti-TNF drug response in AIDs.

When we retrieved data about expression levels of genes with missSNPs, or AID risk genes, the obtained expression pattern indicated their high expression in several immune-related human tissues and cells. The same genes were also expressed in several adult tissues, but also in fetal tissues. Few genes, like MICA and IL13 did not have noticeable expression in any tested tissue; others, like IL6R and NOD2 were expressed very scarcely. Expression pattern is very important, because obviously, only genes that are expressed simultaneously may interact and form networks or pathways.

Figure 6. Expression pattern of AID missSNP genes

The Cytoscape network analyses enabled inclusion of various gene data, more inclusive than other tools or databases. At the same time, that is exactly why network analyses by Cytoscape have the capability to make imputations based on broad data sources. Tools that are more conservative and databases like STRING, ConsensusPathDB and KEGG cannot execute the same analyses and none of the Cytoscape results can be found by using these tools.

3.3.1.2. Network analyses using STRING software tools (STRING network analyses)

We also analysed networks of GWAS AID SNP harboring genes and TNF constructed by STRING software tool. Though the STRING tool does not have the same potential for intersecting various networks as Cytoscape does, it allows for evaluation of significance of protein-protein interactions (PPI) for the analysed set of proteins.

The STRING network images in the Figures 7a, 7b, and 7c present missSNP, ncSNP and allSNP network respectively. All networks were constructed with the high confidence (index of 0.9) and were significantly enriched in protein-protein interactions (PPI) with p values lower than $1x10^{-4}$. The enrichment level p values are shown separately in the Table 10. Most of the interactions were identified with the highest confidence from text mining data, but also from biochemical experimental data and coexpression data.

Figure 7. Image of networks constructed by STRING for missSNP, ncSNP and allSNP datasets

7a. missSNP network

7b. ncSNP network

7c. allSNP network

The edges among missSNP harboring proteins represented textmined connections, but they were also enforced by biochemical and coexpression experimental data. The edges indicated only 3 interaction among the missSNP set, between TYK2 and IL23R and IL6R. NOD2 and ATG16L1 were linked by edges indicating textmining and coexpression connections. However, TNF was not detected to interact with any of missSNP proteins using STRING. In case of the ncSNP set, eight interactions were observed with the highest confidence (index of 0.9). They also represented text mining data, mainly enforced with experimental data. IRF1 and IRF5 had edges detected by three methods: textmining, co-expression and presence in curated databases. The two chemokines' receptors CCR6 and CXCR1, had additional binding detected in co-precipitation biochemical experiments. The rest of edges between protein pairs represented textmining data and experimental data. TNF was linked with three proteins, CD40, TRAF1 and IL10, the later being only text mining based link. For the compiled set of allSNP dataset that contains missSNP set and ncSNP set together, several new links were detected among genes/protein nodes, which were not recognized before in the missSNP set, nor in the ncSNP set. In the allSNP set several nodes were connected with edges representing experimental data for following pairs: IL10 and IL6R, PTPN22 and CTLA 4, IL10 and TYK2, NOTCH4 and RBPJ. The links based on textmining, co-expression and experimental results were detected between TNFAIP3, TRAF1, CD40 and TNF, connecting for the first time missSNP proteins TNFAIP3, and IL6R, IL23R and TYK2 on one side and NOD2 and ATG16L1 on other side. Only when the allSNP dataset was tested by STRING, we were able to confirm partially the results detected by Cytoscape network analyses.

These three genesets were further tested for PPI enrichment (Table 10). STRING results indicated that the all three sets were highly enriched for PPI, as the number of detected interactions (edges) outnumbered number of expected interactions, even by several orders in case of the missSNP set.

Table 10. PPI enrichment for missSNP, ncSNP and allSNP genesets computed by STRING

	missSNPs set	ncSNPs set	all SNPs set
p value:	9.1 e-4	6.3 e-4	1.8 e-7
interactions observed:	3	8	16
Interactions expected:	0	2	3
number of proteins:	24	40	61

The fact that ncSNP set was enriched for PPI, opposes a possibility that the set was not enough coherent. When two sets had been analyzed together, the allSNP set was also highly enriched for PPI, exceeding the sum of PPI for both sets. This finding indicates that ncSNP set brought some PPI enrichment to complement the missSNP set, confirming that our selection of genes performed at the level of SNP-gene prioritization for ncSNPs was the right one.

Network analyses by STRING have shown that the SNP sets are enriched for PPI. It also indicated several links among SNP harboring/influenced genes as well as with TNF.

3.3.2. Pathway analyses

3.3.2.1. Identification of KEGG pathway dataset for the GWAS AID SNP genes

To clarify and confirm the encouraging findings from the Cytoscape network analyses of AID SNP genes, as well as some STRING data, which found common interactants between several AID GWAS missSNP genes and TNF, we further searched the KEGG database for all pathways that might home (contain) the selected coding genes harboring AID SNPs or being influenced by AID SNPs. After we found KEGG pathways (AID SNP pathways), we analysed their intersections for the common genes/proteins between two KEGG pathways.

3.3.2.1.1. Identification of KEGG pathway dataset for the GWAS AID missSNP genes

More than 30 KEGG pathways that contain GWAS AID missense SNP harboring genes/proteins have been identified and retrieved for analyses (Table 11). The retrieved KEGG pathways (missSNP KEGG pathways were only selected for the experiment) are associated with one or several AIDs via the inflicted genes. Some were even common to other immune system diseases.

Seven of the missSNP harboring genes (out of 23 missSNP genes) did not have currently any annotated pathway in the KEGG database (nor in any other pathway database including NCI, Reactome, InAct, etc.; data not presented). In addition, the same genes lacking pathways, were often poorly annotated elsewhere (NCBI Entrez Gene or European Ensemble databases or Havana Project), and even their basic functions were almost always unknown.

Table 11. KEGG pathways for the genes harboring missense SNPs

There was no obvious relationship between the number of pathways in which the missSNP gene participate and the number of associated diseases. Much "wired" genes, like TYK2 (non-receptor tyrosine-protein kinase), participate in numerous pathways and are associated with several AIDs. Others, like FCGR2A (receptor for Fc fragment of IgG) or NFKBIE (NF-kappa-B inhibitor epsilon), are also very "wired" by their presence in several pathways, but are associated with only one or two AID diseases. In addition, several missSNP genes had only one participating pathway, like PTPN22 (protein tyrosine phosphatase, non-receptor type 22 (lymphoid)), but that single pathway is present in almost all AID phenotypes (and in other immune/inflammatory diseases, like SLE or T1D).

3.3.2.1.2. Identification of KEGG pathway dataset for the GWAS AID ncSNP genes

The ncSNP genes dataset was also tested for participation in KEGG pathways after exclusion of HLA genes, RNA genes and pseudogenes (Supplemental Table 3 and Table 12). The RNA genes were excluded because they lack any established involvement in the known KEGG pathways as stated before. Pseudogenes such as ncSNP genes USP81P or WASF5P are generally not translated, even not transcribed and their role in pathways, although suspected, is far from being clear. However, they continue to be very intriguing participants, as they might have roles in the immune system. The HLA genes were excluded because of the complexity of their genetic structure. Although the HLA participating pathways are well known, they were not included in our study, because they all function only in one biological process, antigen peptides presentation, and for that reason have significant, but limited consequence on pathogenesis of AID.

When it comes to existing annotations of genes, the same that was valid for the missSNP genes, was valid for the ncSNPs genes: generally their annotations are very poor. The functional properties are even less known for ncSNP genes than for missSNP genes, and they are less integrated into the KEGG pathways than the missSNP genes (Table 12).

We evaluated ncSNP genes not only for their potential to participate in the KEGG pathways, but also for their functions related to associations with other immune genes, or their relevance in the AIDs and/or other inflammatory/ immune related diseases (Supplemental Table 3).

These additional data for all ncSNP genes were collected from NCBI Entrez Gene, KEGG and the NHGRI Catalog. Only genes that currently have any rationale backed by existing data and by the established criteria for implication of a role in the immune system network were taken into consideration for gene-pathway prioritization and they were considered ncSNP gene set (Supplemental Table 3).

Supplemental Table 3. AID GWAS ncSNPs influenced genes and their participating KEGG pathways

Significant part of ncSNP genes (the genes we found to be influenced by or aligned with ncSNPs) had no known function. However, some of the ncSNP genes do participate in the KEGG pathways.

Majority of these ncSNP associated KEGG pathways have been already presented among the missSNP gene set pathways (and not only because of scale relationship between genes and pathways, since as many as several dozens of genes/proteins might participate in only one pathway). Other ncSNP pathways were not previously found among missSNP gene dataset of pathways (Table 12).

Table 12. KEGG pathways of the genes harboring non-coding SNPs

The ncSNP KEGG pathways, not previously detected in missSNP gene KEGG set pathways, are Intestinal immune network for IgA production and Toll-like receptor signaling pathway, both linked to CD40 gene. Other pathways linked to CD40 gene, a gene known to be involved in the immune system, are T cell receptor signaling pathway and Jak-STAT signaling pathway. These pathways are the key signaling pathways already detected amongst missSNP KEGG pathways.

Our finding reflects a concept that it might be insufficient to take into consideration only missSNP variants from AID GWAS in order to find underlying mechanisms. The detected missSNPs and ncSNPs genes KEGG pathways (we named them all AID SNP KEGG pathways) might be all relevant for SNP harboring genes functioning; we defined all of them as potentially pathological pathways for the AID diseases under study.

3.3.2.1.3. Classification of AID SNP KEGG pathway into modules

We further classified the AID SNP genes containing pathways according to the pathway KEGG classification as shown in the Tables 13 and 14. The pathway classification is important step, because it enhanced our understanding of more advanced modules in which the pathways are thought to operate. However, the modules classification is still a work in progress and many pathways are still only partially classified into the modules (according to the KEGG database).

Table 13. Classification of missSNP KEGG pathways

Table 14. Classification of ncSNP KEGG pathways

More than half of the detected pathways represent description of pathogenesis of human diseases, either systemic or infectious diseases (Tables 13 and 14). These disease pathways are very complex pathways consisting of assemblies or pathways, and they usually encompass several basic signaling and developmental or system bound pathways within each disease pathway.

Apart from these disease pathways, there is much smaller number of basic or core KEGG pathways (Table 13 and 14). We found only 21 basic pathways among the missSNP KEGG pathways and 15 ncSNP KEGG pathways. We consider all listed basic pathways, both missSNP KEGG and ncSNP KEGG to be pathogenic pathways in AIDs. Among these basic pathways, only 6-7 pathways belong to the signaling pathways. Interesting enough, majority of GWAS AID SNP impacted genes belong to these pathways: more than twelve SNP genes participate in only 5 core pathways (data not shown).

Since our research was aimed to find common, intersecting genes between curative and pathogenic pathways, only the SNP KEGG pathways present at the same time in the dataset of TNF pathways were further taken into consideration.

In order to find the overlapping, common pathways, we had first to define TNF pathways, which aTNF drugs use once administered into human body.

3.3.2.2. Identification of KEGG pathways of five anti-TNF biologics and TNF

In order to find potential overlapping of pathways between AID SNP gene pathways (defined as pathological AID pathways) and anti-TNF biologics (curative) pathways, the next step was to identify the pathways that are engaged by the five anti-TNF biologics drugs used for AID therapy. The pathways of

five anti-TNF biologics are annotated in the KEGG pathway database, retrieved and presented in the Table 15:

Table 15. Anti-TNF biologics drug pathways

Only four pathways were shared between all five drugs/biologics (highlighted) and a few others are common to the majority of aTNF biologics. The reason for different usage of the pathways by the aTNF biologics is unclear, as four out of five drugs have relatively similar targets, and only slightly different protein structures (data from Drug Bank). The differences in the existing pathway usage for aTNF drugs a provided in the Table 15., do not coincide with the differences in the molecular structure of the five aTNF biologics and cannot be explained other than the incomplete data.

The aTNF biologics have more than 12 targeted proteins other than TNF (TNF is a biointeractor with the bioactive role) with no known pharmacological action: all subunits of the complement component 1(C1 q, r, and s), Fc receptors with high or low affinity (Fc receptors) and Prostaglandin G/H synthase 2 (data collected from Drug Bank).

The reported narrow pathway usage found in the KEGG databases for aTNFs does not reflect their targets. We considered that it would be erroneous to use only the reported anti-TNF pathways, and neglect all known TNF pathways through which the aTNF biologics might act as well. We concluded that it was reasonable to assume that all currently known TNF pathways have been influenced with the a-TNF biologics therapy, and not only a few annotated. Table 16. provides all known canonical TNF pathways (as retrieved from KEGG database). They all are able to propagate anti-TNF drugs' action and for that reason, we considered all TNF pathways important for the research:

Table 16. KEGG TNF pathways

We analysed the position of TNF in every pathway and made conclusions regarding whether the TNF is a signaling molecule in a pathway, or it is a product of a pathway. This is an important difference, because the direct effect of aTNFs as inhibiting molecules is binding to TNF and inhibiting TNF action in the intercellular space. Many of these KEGG TNF pathways only claim that the TNF is its participant, but do not provide clear role assigned to TNF as its member (Table 16.). Among the pathways, only few pathways have TNF as a signaling, trigger molecule (12 pathways out of 45 TNF pathways), or as an end-product (9 pathways out of 45 TNF pathways), and two pathways have TNF both as a signaling molecule, and as its product. More data would be needed to provide clearer role of TNF. However, all listed TNF pathways from KEGG database were further used to search for overlapping pathways

between GWAS AID SNP pathways and TNF pathways, although our analysis suggested that not all pathways might contribute equally to the TNF role overall.

3.3.2.3. Curative pathways: KEGG pathway dataset common to both GWAS AID SNP pathways and TNF signaling pathways

Not all SNP gene pathways were influenced by TNF participation; only ones influenced by TNF were considered further for analyses. We cross-examined both sets of pathways and detected GWAS SNP KEGG pathways that are shared with known TNF pathways. The ncSNP gene pathways and missSNP gene pathways were separately analyzed for common members with TNF pathways and the resulting subsets of overlapping pathways are presented in the Supplemental Table 4 and Supplemental Table 9 (color coded for easier identification).

Supplemental Table 4. Comparison between all pathways: missSNP gene pathways, ncSNP gene pathways and TNF containing pathways

We found that three pathogenic AID SNP pathways, common to both missSNP and ncSNP pathways, did not have any members common to TNF signaling pathways: two important signaling pathways, Notch signaling pathway and Jak-STAT signaling pathway, and one immune system pathway, Intestinal immune network for IgA production. Consequently, these three pathogenic pathways were not considered further as potential curative (intervention) pathways, because simply they could not be influenced by anti-TNF therapy (Table 17). Similarly, B cell receptor signaling pathway, a very important missSNP KEGG pathogenic pathway, is not influenced with TNF at any point (Table 17).

Only pathways that are common to the group of TNF pathways and the group of SNP gene pathways we consider being important paths for the anti-TNF biologics action (intervention capable). We named these KEGG pathways potentially curative AID pathways and presented them in the Table 17, together with pathways that cannot support TNF action because they do not contain it, but may still influence pathogenesis of the AID.

Table 17. Relationship between AID SNP pathways and TNF pathways

All three pathways selected for ncSNP genes associations have weaker links than missSNP harboring genes. MAPK signaling pathway was extracted based on gene GNA12, a gene that has very few annotations and is not highly probable to participate in the pathways. The same is valid for gene SMAD3 that led us to TGF-beta signaling pathway. Genes CD40 and IRF1, IRF5 are linked with Toll-like

receptor signaling pathway and are well-annotated genes, but the pathway is very complex and insufficiently linked to TNF (TNF is only a product of this pathway: Table 16).

Hematopoietic cell lineage pathway has very uncertain role because it is a very complex pathway, a collection of huge number of proteins and role of IL13 gene liking AID SNPs to it is unclear. Adipocytokine signaling pathway, linked to NFKBIE gene, is not clearly annotated as linked to immune system, and its potential role is clearer in T1D, than in any of AID under study.

We wanted to confirm these results from the KEGG database using Cytoscape tools. In order to find intersecting members (genes/proteins) of the selected SNP AID pathways and TNF pathways, we used the Cytoscape to query intersections of these pathways retrieved from KEGG database. For this interrogation, SNP and TNF KEGG pathways were queried among themselves in pairs. The following results were found in the Supplemental Table 5.

Supplemental Table 5. Subset of pathways common to AID missSNP pathways and TNF signaling pathways: intersections analysed by Cytoscape

As concluded from this experiment, only few SNP AID pathways contain TNF signaling pathway members in the intersecting gene sets. The data confirmed previous results shown in the Table 17 for majority of the TNF-SNP common pathways.

It is very important to stress that anti-TNF blockers (anti-TNF biologics) do not influence the SNP AID pathways that did not contain the intersecting set.

For example, the NOTCH signaling pathway is completely independent from TNF signaling and from all other SNP-gene-harboring pathways. Previously we discovered NOTCH pathway as a potential pathogenic AID pathway; it turns to be one of SNP pathways not able to “succumb” to TNF modification, not able to interfere with TNF signaling. For that reason, we think that anti-TNF biologics are not able to modify the NOTCH pathway.

However, TNF could influence Osteoclast differentiation pathway, at least a mayor part of this pathway. TNF also could influence B cell receptor signaling and Ubiquitin mediated proteolysis according to Cytoscape analyses, contradictory to our earlier results for these pathways obtained from KEGG database.

Several other listed pathways could be linked via TNF and “talk” among themselves only because of the presence of TNF gene: NOD like signaling, Antigen processing and presentation, and Cytokine-cytokine

receptor interactions (actually not a real pathway, but more collections of receptor-ligand complexes). These three pathways are examples of the pathways influenced by TNF. They could be modified by anti-TNF biologics.

The Jak-STAT pathway could influence only cytokines and cytokines receptor containing pathways; there are quite a number of missSNP genes in this group, among cytokines and cytokines receptors (IL6R, IL13, IL7R, Il17REL, etc.). However, Jak-STAT pathway had no influence on TCR and BCR signaling, nor on the Antigen processing and presentation, IgA production in intestines, or on NFkB signaling pathway. The TGF-beta signaling and MAPK signaling pathways, found in the ncSNP pathway group, also had no common members with Jak-STAT pathway (data not shown in the tables).

In conclusion, the results in the Table 17 and Supplemental table 5 showed separation between at least two groups (if not more) of AID GWAS pathways that could communicate intra-group, among themselves, but not inter-groups, between the two groups. If we consider the AID GWAS SNP gene pathways potentially pathological pathways, then these two pathway groups (or more) obviously have consequences on the therapy with anti-TNF biologics. The image illustrating at least two separate groups of pathways was shown in the Figure 8.

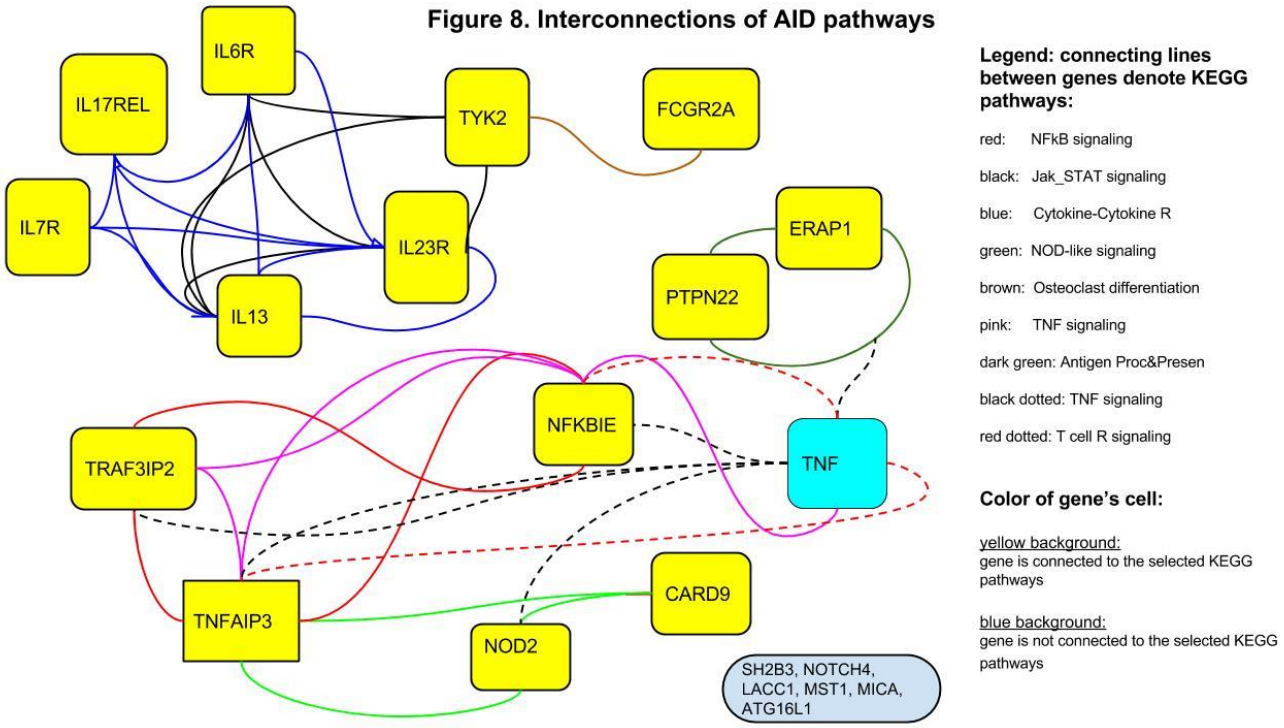
In this illustration, missense SNP genes are shown as the nodes connected with the edges representing KEGG pathways. Two distinct groups of SNP harboring genes are clearly separated as they do communicate intra groups via different pathways.

Several missSNPs harboring genes (SH2B3, NOTCH4, LACC1, MST1, MICA, and ATG16L1) do not belong to any of the two well-defined groups of pathways. NOTCH4 participates only in its own pathway, NOTCH signaling, which is completely separated from any other pathway. The similar case is with SH2B3 that only belongs in the Neurotrophin signaling pathway, exclusively occurring in neurons.

Figure 8. Interconnection between missense SNP harboring gene pathways and TNF

ATG16L1 only belongs to the pathway for Regulation of autophagy, which has no tissue restrictions, but on the distal end, this pathway interacts with Apoptosis and Antigen processing and presentation pathways. Both pathways have no connections with Jak-STAT pathway or the group of pathways linked to Jak-STAT pathways. Genes LACC1, MST1 and MICA do not have any known KEGG pathway (or any other pathway from other databases; results not presented).

PTPN22 and ERAP1 are part of Antigen processing and presentation pathway. This pathway is influenced directly by TNF, so both genes are connected with TNF signaling pathways.



The Apoptosis pathway is influenced by TNF and NFkB pathway (which is also influenced by TNF). It has several SNP gene members, but has no connections with Jak-STAT pathway.

All pathways with TNF influenced members could be clearly modified by anti-TNF biologics, while others could not. The obtained results are manually extracted first and then confirmed by Cytoscape software tools.

However, for additional confirmation (since we had a few discrepancies, or were not able to test the difference between Cytoscape data and manual KEGG retrieval), we also tested the AID SNP influenced genes and TNF gene pathways with additional bioinformatics tools: STRING and ConsensusPathDB, using gene-pathway prioritization tools based on PPI. The obtained results of gene-pathway prioritization are presented in the Tables 19, 20 and 21.

3.3.2.4. Pathways enrichment for AID GWAS SNP gene sets by STRING

The STRING bioinformatics tool was employed to perform pathways enrichment analyses of the AID GWAS SNP affected genes. Three genes sets were separately analyzed and compared: missSNP gene set, ncSNP gene set and the set consisted of all GWAS SNP genes.

All three sets, (missSNP set, ncSNP set and allSNP set), analyzed for STRING pathways enrichment were determined earlier (Table 8). The scoring level of confidence for pathway enrichment was set at 0.7 or high for all attributes. The STRING retrieved the following pathways:

Table 18. STRING pathways enrichment data for AID GWAS SNP gene/protein datasets

Four core and two disease pathways were enriched for missSNP gene set. However, ncSNP gene set has shown only one pathway, which is more collection of cytokines and their receptors than a directional pathway.

The allSNP set included wider list of KEGG pathways enrichment, than the missSNP set. It showed that the ncSNP set, although was enriched only in one pathway, contributed to the enrichment of its KEGG pathways when assembled together with the missSNP set. The number of genes participating in the enriched KEGG pathways had increased, and new KEGG pathways were added to the list of pathways enriched for all AID SNPs. This might bring us to the similar conclusion as before, that although the ncSNP set was not enriched per se, when added to the missSNP set they represented a unified set of genes/proteins enriched for specific biological functions. All STRING pathway data were considered after the correction for false discovery rate (FDR) as pathogenic pathways for AIDs.

3.3.2.5. Pathways enrichment of AID GWAS SNP genes by ConsensusPathDB

Using ConsensusPathDB the analyses for over-representation of pathways were performed for the same sets of AID GWAS SNPs: missSNP, ncSNP and allSNP sets.

Supplemental Table 6. ConsensusPathDB: enriched pathway sets for missSNP dataset

The missSNP set retrieved 78 enriched pathways from the several pathway databases that serve as source databases for ConsensusPathDB software (Supplemental Table 6). This set appeared in several new pathways not obtained previously from STRING enrichment results, but were recognized before in the data obtained in the Cytoscape analyses. New pathways were NFkB signaling, TNF signaling and Toll like receptors signaling pathways. ConsensusPathDB also retrieved several other pathways that did

not exist in the KEGG, such as IL2, IL4, IL6 and IL23 signaling pathways, though they could be considered as the parts of the more integrated KEGG pathway for cytokines and cytokine receptors.

When the same set of missSNPs was tested for overrepresentation only in KEGG pathway database, only 23 pathways from KEGG were detected after correction for FDR (Table 19).

Table 19. ConsensusPathDB: enriched KEGG pathway-based sets for missSNP geneset

The number of pathways was much higher, indicated better retrieval enrichment of the used tool. ConsensusPathDB was able even in KEGG database to retrieve more pathways than STRING, where only five KEGG pathways were detected.

Pathway enrichment analyses retrieved 26 pathways for the ncSNP set from the several pathway databases used by ConsensusPathDB retrieval tools (Supplemental Table 7).This set showed new pathways, not obtained previously from STRING for the same ncSNP set: T cell receptor signaling and Chemokine signaling pathways; all these pathways came from the KEGG database.

Supplemental Table 7. ConsensusPathDB: enriched pathway sets for ncSNP geneset

Interestingly enough, the ncSNP set pathway enrichment was not indicative of immune system pathways, because only couple of pathways were linked to the immune system, like Cytokine-Cytokine Receptor and T cell signaling pathways (pathways that did not show up in the pathway enrichment analyses for missSNP set).

The allSNP set was tested separately (Supplemental Table 8) for pathway enrichment from all pathway databases sources and from KEGG database (Table 20).

Supplemental Table 8. ConsensusPathDB: enriched pathway sets for allSNP geneset

More than 70% of the allSNP set genes were present at least in one pathway. Huge number of enriched pathway sets was retrieved (85), but most of them were overlapping. The set with the most members was Jak-STAT set and Cytokine-cytokine receptor interactions set with 6-8 members. However, when only KEGG pathways were selected, all seven core pathways were retrieved, in addition to Cytokine-cytokine receptor interactions and Chemokine signaling (both are complex ligand-receptor type of interactions) (Table 20.).

Table 20. ConsensusPathDB: enriched KEGG pathway-based sets for allSNP geneset

The set of seven core KEGG pathways was the same retrieved by the Cytoscape tools. Two additional pathways, Toll-like receptor signaling pathway and Chemokine signaling pathway, were additionally discovered as enriched pathways for the allSNP dataset. This result also confirms our previously shown result after manually searching KEGG pathway database.

This way, we confirmed with other independent software tools the same findings presented in the Figure 6.

3.4. Functional analyses of AID GWAS SNP data using Gene Ontology

In order to understand potential function attributes and connection between AID SNPs and disease mechanisms, we applied GO term enrichment analyses. Gene ontology terms enrichment analyses are independent of PPI and allow extracting knowledge about biological processes (BP), molecular function (MF) and cellular (CC) activity location for the genes of interest. GO term enrichment analyses complement our search for pathways and networks of the proteins that harbor AID SNPs or proteins/genes that are influenced by AID SNPs.

We used STRING and ConsensusPathDB tools to retrieve and analyze GO term set enriched for BP, MF and CC ontology terms assigned to SNP genes. The results are presented in the Tables 21-24. Only the significantly ranked GO terms for the sets are provided, based on p and q values. The numbers and percentages of SNP proteins in the analyzed sets are given, along with the total number of known proteins in a group. We used both STRING and ConsensusPathDB because they provided slightly different data resulting from different algorithm applications.

3.4.1. GO term enrichments by STRING

The STRING enrichment analyses were limited by the number of members in the analytical sets, because the sets had to contain at least ten members in order to be evaluated by STRING. It worked well for the three AID SNPs sets, but finer parsing down to subsets was not possible.

Table 21. STRING BP, MF and CC GO terms enrichment for AID missSNP, ncSNP and allSNP genesets

GO annotations for biological process (BP) indicate that all three sets exhibited GO term enrichment. The GO terms indicated that the sets were enriched for immune system and its function, development and regulation. The allSNP set was the richest in the BP terms as expected (Table 21). However, very

few annotation terms or none existed for terms delineating the molecular function (MF) or cell component (CC) spaces (Table 21).

The most probable reason for these results was that GO terms are biased towards BP terms as MF and CC terms were poorly annotated for the sets (or in general). From the BP terms, it was clear that the SNP sets indicated engagement in the immune system, immune system development, its function and control. Poorly annotated terms for MF and CC were not sufficient to make any conclusions, neither about specific functions nor about location of the analyzed SNP sets. It was known from other data that most of proteins were membrane receptors, lymphokines, enzymes or translation factors, but that could not be confirmed by their enrichment in GO terms.

For that reason, we also employed ConsensusPathDB and its tools that did not have the limitations for enrichment analyses; in addition, the ConsensusPathDB has a richer base of pathways' search space.

3.4.2. GO term enrichments by ConsensusPathDB

Using ConsensusPathDB, the analyses for over-representation of GO terms were performed for the same sets of AID GWAS SNPs: missSNP, ncSNP and allSNP sets (Tables 22-24).

Table 22. Enrichment of BP, MF and CC GO terms for missSNP geneset

GO BP term enrichment indicated involvement in the immune system functions, activation of its components often in response to stress (the most numerous terms) and the development and regulation of the immune system. These findings were in line with previously obtained GO terms from STRING, but they were better classified for GO levels and provided clearer answer to biological processes for missSNP set. Regarding MF and CC, the missSNP set was enriched for terms indicating protein binding, to the less extent antigen binding, and its location was tied to cellular membranes.

More detailed analyses of MF terms enriched for missSNP dataset indicated that the protein binding function was linked to terms describing receptor-ligand binding for cytokines or growth factors facilitated by CARD domain of the interacting proteins. They also included antigen binding function, which is the one of the core of immune functions (Table 22).

Further deeper search for CC terms only returned location descriptions linked to cell membrane, cell surface or its periphery. Higher-level CC terms, like cell organelles, were not enriched for the missSNP set (Table 22).

However, again we have seen the bias towards GO BP annotations vs CC or MF annotations in ConsensusPathDB similar to STRING, except the later categories were less rich and less detailed. It was not clear whether it was caused by the nature of SNP sets or general deficiencies of annotations of GO CC and MF spaces.

The ncSNP set retrieved fewer results, but still was able to indicated clearer picture for their functional engagement like immune response and signaling (Table 23):

Table 23. Enrichment of BP, MF and CC GO terms for ncSNP geneset

GO enriched term sets for ncSNP dataset pointed towards immune system involvement and its regulation and response. They also included terms describing signal transduction and cell migration. Cell component localization was the same as for the missSNP set, but the MF terms indicated signal transduction, protein binding and transcription factor binding for specific DNA sequences.

When we tested the allSNP set for GO term enrichment, we obtain very similar results regarding immune system processes. The most numerous enriched ontology terms indicated immune response and stimulus response, signal transduction, regulation and development. At the MF level, terms were enriched for protein binding and signal transducing. At the level CC enrichment of location terms argued for cell membranes at the periphery of cells as a place of activity of the majority of proteins harboring/influenced by the AID SNPs (Table 23).

Table 24. Enrichment of BP, MF and CC GO terms for allSNP geneset

Almost all biological process terms fitted into known immune system processes (Table 24). The MF terms were connected with protein binding, transcription factors binding and signal transduction, all in line with roles playing in the immune system (or any other system for that matter) (Table 24). The CC terms for allSNP set did not differ from the CC terms for the missSNP and ncSNP sets (Table 24).

3.5. Disease prioritization

3.5.1. Disease Connect DB enrichment for AID SNP dataset

We employed the DiseaseConnectDB bioinformatics database and tools to search from an opposite side of our already presented approach and to double check whether this relatively new tool would confirm our results using other tools (ref).

When the AID diseases were searched for enrichment in the pathway set employing DisseaseConnectDB software tool, we obtained the results very similar about the pathways engaged in the AID diseases to what the STRING and ConsensusPathDB tools were able to provide (Table 25). As shown in the table, all pathways found to be active in AID diseases, indicate pleiotropy on pathways level. We already have shown pleiotropy at the gene level using different bioinformatics tools (Table 3).

Table 25. Disease Connect DB pathway dataset for AID

Out of more than 40 pathways linked to the AIDs by DiseaseConnectDB tools, we selected 20 pathways based on their presence in two or more AID diseases. We believe that even pathways that are currently detected to be present in only two AIDs, mainly CD and RA, will be found more often in three other AIDs once more meta-GWAS are performed and more data is collected for these diseases.

Among the associated pathways, only seven pathways are real core signaling pathways, transducing signals from/to environment and cells/tissues. These pathways are a part of general environmental information processing or signals transduction within the immune organismal system. There are three core-signaling pathways, and four immune systems signaling pathways.

The other 2/3 of overlapping pathways between AIDs belong to human diseases modules, diseases like metabolic, immune or infectious diseases caused by parasites, viruses or bacteria. The DiseaseConnectDB discovered pathways for the AID SNPs, belong to the category of human diseases identical to the pathways prioritized for the same AID SNP genes manually using KEGG database, or by STRING or ConsensusPathDB. They are complex pathways that consist of several core pathways, including signaling pathways and immune system pathways among them.

3.5.2. STRING disease enrichment for AID GWAS SNP dataset

We wanted also to compare these pathways-diseases prioritization results with the disease enrichment data recently available in STRING software system.

Table 26. STRING disease enrichment data for AID GWAS SNP genesets

Analysis using STRING tool for disease enrichment has shown that GWAS SNP genes were significantly enriched (after FDR correction) for the almost all of the seven autoimmune/ inflammatory diseases under study except ulcerative colitis, or all AIDs except psoriasis.

After checking for the complete list of diseases available in STRING database, we realized that UC was not listed in the STRING database. The reason is not clear, but IBD and Crohn disease, are sometimes

considered two separate diseases within UC umbrella; both IBD and CD are on the top of the list of diseases in which the allSNP set is enriched. Again, the ncSNP set did not show enrichment for any diseases, but also had no negative effect on the list of enriched diseases for missSNP set that was identical to allSNP set; the ncSNP gene set was neutral (data not shown in the Table 26).

4. DISCUSSION

There is a strong consensus that genome-wide association studies (GWAS) have allowed acquiring enormous knowledge about genetics of human autoimmune diseases (AID), probably more than any other scientific approach before (Zerhouni and Nabel 2008; Visscher et al., 2012). The agreement comes even from the scientists who deny CD-CV postulate (McClellan and King 2010). The AID GWAS have characterized hundreds of genetic associations as potential risk factors for autoimmune disease development and inheritance. In addition to identification of the inheritance and risk genome structures, the GWAS acquired knowledge has a potential to clarify, even discover novel disease mechanisms by defining the operational AID biological pathways; it also has a potential to pave a road to development of new drugs for management of autoimmune diseases.

Starting from GWAS data, we launched on a task to find why the already approved aTNF therapy for several AID does not work in a good portion of AID patients. One possibility, apart from all others (natural antibody inhibition, acquired immunization to aTNFs, imbalance between pharmacokinetic parameters, under-dosing, etc.), is that the genetic makeup is accountable for the unresponsiveness. There is a possibility that patients who respond and who do not respond to aTNFs, differ in a couple or dozens of gene alleles that are key to aTNFs response. Variability of those gene alleles could be tested ahead of the aTNF therapy and used to make prediction about the responsiveness. Similar situation has been discovered already for several drugs, where an advance allele gene testing helps making suitable therapy choices (Limdi and Veenstra 2008). This thinking represents a base of pharmacogenomics and it is in line with its goals: taking into consideration genome characteristics of an individual, provide the optimal medication that will not trigger negative adverse events, but deliver a cure.

Our rationale was that, by using AID GWAS-generated knowledge, we could define genes and consequently, more important, the pathways engaged in the AID pathogenesis and aTNFs response. We assumed that the pathological AID pathways are all pathways discovered starting from AID GWAS SNPs. We reasoned that a crossroad between the pathways engaged in aTNFs action (which are known), and the pathways responsible for diseases activity (unknown but believed to exist), would point towards genes/pathways responsible for the aTNFs response in AID. Overlapping between the AID pathological pathways and aTNF interventional pathways could define a subset of curative pathways, which should contain key structures responsible for both activities: aTNF actions driving to remission or AID natural disease progression. A successful aTNF therapy induces a remission in AID patients, a condition in which we do not expect the disease pathological pathways to be any more active: a

successful aTNF therapy shuts down the pathological pathways; but when aTNF therapy is unsuccessful, the pathological pathways stay operational and disease remains active.

Incoming new studies, if funded, should study further the polymorphism of these newly defined key genes or pathways in individuals in order to confirm their different functionality in cells and tissues of responders and non-responders. Furthermore, those genes could be studied by combining the information in AID patients' health records with detailed dense genotyping or better, narrow scoped sequencing of the regions, whole exomes (WES) or genomes (WGS) by the next generation sequencing methods (Kiezun et al., 2012).

4.1. GWAS AID Associations

In order to provide insights into the AID disease biology, we encountered several challenges in the interpretation of GWAS association. They had to be first resolved, before we could identify candidate gene sets based on currently available AID GWAS data.

4.1.1. P value

The common practice has been to consider only GWAS associations with statistical significance labeled with p values equal or smaller than $p<5 \times 10^{-8}$ (Manolio et al., 2009). Lower levels of significance may be used, with p values of $p<1 \times 10^{-5}$, even $p<1 \times 10^{-4}$, in order to include more associations and still obtain meaningful results in SNP- gene prioritizations. However, such high p values, used as a cut-off p values, might condition retrieval of many false positive associations. Too many false positives might "dilute" SNP-gene prioritization results and result in no enrichment in parameters that define gene set functions. No enrichment might originate from the fact that selected genes have no common function. No enrichment might result from the ratio between genes with coherent functions vs. genes with incoherent functions; in that case, the last group obscure the first group in the same gene set. On the other side, if the cut-off value is too low ($p<5 \times 10^{-8}$) or more, many legitimate associations are lost and with them a part of real biological functions (Visscher et al., 2012). Some researchers argue that even associations with $p=5 \times 10^{-3}$ are meaningful (Visscher et al., 2012). After testing a few sets for results, we decided to crop all AID associations with p values equal or smaller than 1×10^{-7}. However, in our case, very few AID associations had p value of 1×10^{-7}, as the majority of AID SNPs had much lower p values (Supplemental Table 1). The selection of SNP association cut-off p value was especially important, because some AID have few GWAS associations, yet we wanted to include associations for every AID in which aTNF therapy has been applied (an example is AS with very few associations found so far). However, unless other relevant orthogonal data supported a SNP with a p value of 1×10^{-7}, or higher, we avoided using it for gene or

pathway prioritization analyses. NHGRI Catalog collects all association with p values equal or lower than $p<1 \times 10^{-5}$ as real and justifiable associations, keeping as many associations as possible.

We retrieved associations for each AID under study and treated them as a pool of data of equal value, not giving more weight to a disease with more GWAS SNP or lower p values. We were not interested in differences between AIDs, because for our purpose, the bigger pool of SNPs provided larger searching space for SNP-gene-pathway or network prioritization, since we have been interested in finding common ground why aTNF works/or not in all of selected diseases.

The AID associations used for our research mainly encompass common SNPs with high frequency (MAF ≥ 5%), but also common SNPs with low frequency (0.5% <MAF <5%). We sourced associations only from the NHGRI Catalog, because NHGRI associations are curated (see Methods), while associations from other sources are not. The associations taken directly from publications suffer from lack of the necessary review and selection criteria. We did not use very low MAF SNPs (0.1%< MAF < 0.5%), nor we used SNPs available from sources other than NHGRI Catalog. We did not collect nor analysed rare SNPs, because they are not in the Catalog, as they are not intentionally recognized by GWAS methodology. However, choosing lower cut off for p values does not necessarily mean elimination of rare variant signals, because there is always a possibility those SNPs with lower p values do reflect rare mutations, possible even several rare mutations grouped in a narrow gene segment. Rare variants are especially indistinguishable if they are in very high or perfect LD, as it is the case for HLA complex.

We also did not use combination of significant SNP associations within a gene in a broad sense in order to obtain lower p values, because combined SNPs could not be analyzed by the software tools we used. If one gene region has two or even several SNPs associated with an AID, it is logical to assume that the gene has more weight overall as a risk factor for the AID. The situation is even more complicated when two SNPs are in high or perfect LD, as they cannot be analyzed for their combined influence, which obviously exists (our examples are two missSNPs in NFKBIE gene in perfect LD). Again, there is no available software that could analyze these cases.

Some recent data claims that pooling GWAS studies with a significant pleiotropy between diseases, allows for a less stringent new statistical method. It enhances discovery of new SNPs from already published studies, which might explain additional hereditary component, by increasing p values to 10^{-5} or even less (Lilley and Wallace 2015). It is important to keep in mind that genes might be incorrectly included in gene sets not only because of p values, but also because of erroneously assigned functions by poor annotations or lack of it.

4.1.2. Rare vs. common (and low frequency common) SNPs

Several common AID missSNP risk factors in our study are not exactly common SNPs by the strict definition, as their minor allele frequency (MAF) values are lower than 5%. However, they are considered common SNPs with lower frequency (0.5 %<MAF<5%) in European population, and not rare variants (MAF<0.1%). The examples of common SNPs with lower MAF are the missSNPs in TYK2 and IL23R genes, which we found to be associated with multiple AID (Table 2). Both missSNP are situated within highly conserved regions, typical for common variants.

The rare risk variants (MAF<0.1%), otherwise undetectable by GWAS, might still be responsible for an association signal in AID detected by GWAS, because rare SNPs might be in high LD with lead SNPs (Beaudoin M, et al., 2013). When analyzing SNPs in high LD with a lead SNP, rare variants could be among them as well, because rare SNPs are present in the dbSNP and therefore used by software tools like the HaploReg. Although we did not collect and analysed rare SNPs, rare SNPs might signal via common SNPs in GWAS that we analyzed, as it was impossible to eliminate rare SNPs. The status of rare SNPs in common disease is still an issue on which the opinions are sharply split (Schork et al., 2009; McClellen and King 2010; Visscher et al., 2012; Hunt et al., 2013; Lee et al., 2014). However, recently it was shown that rare coding-region variants at known AID loci play negligible role in common autoimmune disease susceptibility (Hunt et al., 2013.). With the improved detailed annotations of SNPs from HapMap3 and 1000G projects, more low frequency common SNPs have been linked to GWAS associations. Currently, up to 60% of known SNPs in the human genome are variants with an allele frequency <5%. Low frequency common variants of moderate effects are likely to play an important role in uncovering complex disease/trait heritability that has not been explained so far by high frequency common variants. However, low frequency variants are often population specific: e.g., majority variants are population specific between European Americans and African Americans (Tennessen et al., 2012). Factors such as rapid population growth and weak purifying selection have allowed ancestral populations to accumulate an excess of low frequency variants across the human genomes (Tennessen et al., 2012).

The CD-CV hypothesis vs. the CD-RV hypothesis discords argue that the genetic variability in the human population with considerable frequency and low penetrance contributes to genetic susceptibility to common diseases vs. that the rare genetic variability with high penetrance is the major contributor to genetic susceptibility to common diseases. The current debate is about the frequency of disease-causing alleles. The CD-CV is accepted by many as an explanation of the risk factors for common diseases (Reich and Lander 2001), in spite of widely accepted realization that common risk GWAS variants have failed to explain the vast majority of genetic heritability for any common human disease

(Manolio et al., 2009; Visscher et al., 2012). They are defending the CD-CV theory with arguments that only common variants add up when the extent of common diseases in humans is calculated (Schork et al., 2009). They claim that the statistics used in GWAS is depressing heritability, and are proposing improvements in statistical approach able to explain a large proportion of heritability, claiming that the rest of unexplained heritability is contributed by non-variant causes (Visscher et al., 2012; Moser et al., 2015).

Other side argue that rare variants (defined by sequencing techniques as a MAF<0.1%), are the only alleles that may contribute to complex diseases (McClellan and King 2010). Rare SNPs cannot be neglected because the average ratio of rare to common alleles is 6:1, indicating overwhelming majority of rare genomic variants in the human genome (Tennessen et al., 2012). Since rare genetic variants are predicted to outnumber common variants, it is important to keep in mind they still might provide lead SNP signals observed by GWAS, especially when the lead SNP has no functional effect on disease risk as judged by the orthogonal evidence. The influence of rare variants could be resolved only with sequencing of the whole genomes or exomes, or smaller regions. However, the same daunting problem of finding which rare variants are disease-causing mutations will stay problematic as before; sequencing does not avoid a trap of multiple measuring (McClellan and King 2010; Kiezun et al., 2012).

The nature of the genetic contribution on individual susceptibility to common complex diseases is still unresolved. Much more data will be needed for any argument to prevail. Whether the causing variants are rare or common, nevertheless, they must influence biological processes by modulating them. Knowledge about how mutations disrupt or modulate key biological pathways and networks underlying autoimmune diseases is essential for understanding the biology of diseases and potential drug intervention. The GWAS data have led to that goal if not completely, then at minimum partially, as one way of acquiring necessary knowledge about polymorphism of pathways and potential drug targets.

4.1.3. Genomic context, location, and LD of associated AID SNPs

The second problem we had to resolve was the exact genomic context of the AID SNPs retrieved associations. Occasionally, we had to resolve contradictory information on few locations assigned for a SNP, using the newest data from 1000 G project and latest version of HaploReg3.

Before assigning a genomic context of all retrieved associations, we checked each SNP location in the dbSNP database. Knowing the exact location of an associated SNP and the most recent version of gene or SNP location in the NCBI databases, did help further analysis such as impact of amino-acid change, but it did not necessarily resolve a cause of association signals. Causal variant is an unknown variant

that has direct or indirect functional consequences on disease risk, even if there are no tools to prove causality. Even when the exact location of a SNP was resolved, and its genomic context was defined, the most probable cause of the association signals still had to be analysed with available tools and annotations.

The associated loci (GWAS tag SNPs) usually signal variable size regions, which contain multiple candidate genes for disease susceptibility in LD. They do not provide direct, restricted information about an exact location. This is conditioned by the human genome architecture and nature of inheritance. Limitations of GWAS biochip design, including the 200K SNP Immunochip used for the majority of AID GWAS, also contribute to this locus uncertainty. Since LD varies along the chromosome, a size of an LD segment had to be assigned for each SNP. We used HaploReg-v2 and -v4 software to determine LD for each AID SNP based on the newest updates in annotations.

When the associated lead SNPs are nonsynonymous, missense SNPs, they are by definition considered functional SNPs. However, that does not mean they are the causal SNPs for disease risk factors. It has to be confirmed whether a missSNP is responsible for a functional change in a coded protein, which further has a potential to contribute disease biology. If there is other missense SNPs in the same high LD gene region not tagged by the GWAS, it additionally complicated the conclusion. The outcome is vague, because the tagged missSNP and not tagged missSNP (if not both, when their MAFs are high) can influence the protein function in different ways (Farh et al., 2015). For many SNPs, we still do not know whether they actually have a direct functional effect, even when they belong to the “obvious” candidates, which alter the amino acid sequence of the coded proteins. In addition, a cause of an association signal with a missense SNP might be any SNP in LD, not necessarily other missense SNP in the same gene or a gene in high LD. A lead missense SNP could be in high LD with SNPs in other genes, being it missense SNPs or not, if they are causal variants. Only functional experimental results or prediction models may resolve this issue, but the prediction models are still too complicated (Visscher et al., 2012) and experiments are always expensive.

In our study, we have examples of all above discussed situations (Table 2 and Table 4). We resolved all of them using orthogonal data from NCBI and KEGG databases as presented in the Results. We restricted our choice of the SNPs for gene prioritization, expression regulation, and network and pathway analyses to the very high or perfect LD SNPs, newest version of annotations in NCBI and KEGG databases and by using the most updated Cytoscape, HaploReg and RegulomeDB software tools. However, again, only functional experiments will be able to resolve our predictions.

4.1.4. Why there are no AID SNPs located on sex chromosomes?

We did not detect any GWAS AID SNPs on sex chromosomes with the conditions we used for their retrieval from the NHGRI catalog, although AID are promising case studies for investigating the role of X chromosome in disease. Most AID have higher prevalence in females and loci on X chromosome have been suggested before as potential risk factors for autoimmune diseases (Ober et al., 2008). Several AID are so prevalent in females that they might be considered sexual dimorphic complex traits (Chang et al., 2014). Although AID have been extensively studied by GWAS, very few of these GWAS have studied the contribution of X and, combined, have provided little evidence for its role in determining disease susceptibility. Only 15 X chromosome associations out of the 2,800 significant associations were reported by GWAS in nearly 300 diseases/traits, making barely a half percent of all considered SNPs (Hindorff et al., 2013; Wise et al., 2013). GWAS either omitted X chromosome sequence or erroneously linked the data (Wise et al., 2013). However, the chromosome X constitutes 5% of the nuclear genome and underlies almost 10% of Mendelian disorders registered in Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2009).

Specific attention has been paid to this sex chromosome exclusion problem very recently, by introducing software for carrying out X-wide association studies, which will help finding more AID SNP associations. Several X linked genes were found that influence immune response pathways, and particularly may explain additional heritability of several autoimmune diseases (Chang et al., 2014).

4.1.5. Immunochip and its influence on AID SNPs

Majority of GWAS studies regarding autoimmune/inflammatory diseases have been performed using Immunochip 200K or its upgrades (Illumina Inc., San Diego, CA). The Immunochip was introduced by the Wellcome Trust Case Control Consortium (WTCCC) as a powerful tool for immunogenetics gene mapping with preselected tags, believed at that time to include all parental SNP that may influence immune system function. The Immunochip is a custom-made Illumina array containing nearly 200,000 SNPs mapping to 186 loci previously genetically known to be associated with 12 autoimmune diseases preselected by the WTCCC (WTCCC 2007). Its cost was much lower than comparables, allowing advantageous usage of the same technique across the board.

There are several limitations of the biochip used for majority of AIS GWAS studies. One of the major is that Immunochip is not unbiased biochip platform, because the SNPs are preselected based on the previous knowledge about the immune diseases (both knowledge about and scope of immune diseases have advanced a lot since then). Although the GWAS methodology is unbiased, the tag SNPs selection

is not, and may limit or distort the GWAS outcomes. Furthermore, the selected tag Immunochip SNPs do not cover regions with no previous associations ever found, nor do they cover the whole genome. Immunochip is not able to unearth SNPs with unsuspected links to common immune diseases, which is an undisputed possibility supported by surprising GWAS results (Yu et al., 2011). Immunochip is not covering the whole human exome or genome, not even the reference DNA sequence that is skewed towards European populations. Additional weakness of the Immunochip is that it is biased towards SNPs found in populations of European ancestry (EUR) not necessarily present in other populations. However, newer variants of Immunochip have been changed by including SNPs from other populations, but that cannot twist existing AID GWAS, since in all meta-analyses only SNPs that exist in all studies may be taken into consideration. It is also insufficiently designed for HLA region complexity, although HLA genes undoubtedly make a base of several important immune recognition and response processes.

The selected AID GWAS studies were performed by Immunochip (actually by several generations of immune-biochips). They mainly refer to the EUR populations (to the best of our knowledge) and for that reason, our all analyses using HaploReg have been always referred to the same population. However, differences in interpretation may always be a source of unintentional errors. Especially difficult is to restrict North American studies with significant underlying admixture populations, knowing in particular that LDs in African population are different (much shorter regions) from ASN and EUR populations (Hinch 2011).

In addition, the 200K SNP Immunochip contains a substantial proportion of lower common MAF polymorphic variants. They are responsible for uncertainty of identifying variants that tag regions of variable sizes containing multiple candidate genes for disease susceptibility. The lower MAF strongly associated variants SNPs tend to be in LD with the causal variant, rather than have a biological function themselves (Trynka et al., 2011). Newer upgrades of Immunochip are enriched with SNPs from 1000KG project and HapMap3, producing better outcomes (Miller 2013). Even improved chip designs with up to one million or more SNPs might suffer from the same problem, insufficiently reflecting the complex architecture of the human genome. It is not clear if the randomly chosen tag SNPs, evenly distributed along the chromosomes, would help better detection of relevant SNPs.

4.2. Functional effects of SNP-gene prioritization

Accurate functional annotation of SNP variants to the proper regions of the genome is crucial for understanding the biological significances of GWAS results for the underlying mechanisms of associated diseases (Ward and Kellis 2012). Once we selected variants, removed questionable or high p value and eliminated duplicates, the next step was to find if the selected AID SNPs were connected to known

structures with defined functional role, including genes in high LD. All retrieved AID SNPs were tested for their potential functional and regulatory roles or influence on regulatory processes. For it, we used the publicly available functional annotation software: PolyPhen, HaploReg and Regulome DB.

4.2.1. Functional effects of AID missense SNPs

We found 23 coding SNPs in our study (Tables 2 and 3). Coding variants could contribute to altered gene functions in different ways, depending on the characteristics of variants: non-synonymous variants cause amino acid sequence change in the corresponding coded protein and might induce protein function changes depending of the missSNP location; synonymous variants (cd-syn) may alter the translation efficiency (Kubo et al., 2007).

Functional effects of the missSNPs were studied using PolyPhen-v2 software tool to evaluate effect of the amino acid switches on the function of corresponding proteins. Problem often was that not all missSNPs ware completely and correctly catalogued for their functional evaluation by position, identifier, splicing etc. An incorrect position of a SNP may be triggered by different versions (hypothetical or recently corrected versions) of the coding protein, especially for less annotated proteins, or by difference in its splicing pattern. In addition, PolyPhen software often does not agree with the position or content of amino acid change from the databases; identifier (rs) numbers sometimes are mismatched as well. In search of an acceptable (most accurate) position of missSNP, several databases had to be consulted before the result was obtained. If a missSNP could not be recognized by PolyPhen software tool, it was not evaluated for its functional effect. We had a couple of missSNPs that could not be evaluated by an earlier version of PolyPhen–v2 we used; however, their function was resolved only recently by the upgraded PolyPhen version and with additional help of HaploReg-v4 (examples are missSNPs within boundaries of genes NOD2, IL7R and IL17REL).

Currently there is a debate what represents a modification of a protein conformation caused by an amino acid change. Because proteins are very complex structures, it is possible that the level of change for one domain of protein does not necessarily pertain for the other domain. Less annotated proteins (many SNP harboring proteins/genes have no known function or no annotation, or even no orthologs) do not have well defined active sites, and it is improbable that without that specific knowledge any tool can estimate valid level of functional distortion produced by an amino acid switch. If the active binding site is not defined, (it is usually defined by the known protein function), then it is less probable that the right evaluation of amino acid change could have been achieved. In addition, allosteric modifications occur independently of change in active sites, and influence protein function by changing its normal

interactions. Many proteins have known active sites, but their regulatory sites are still vague (Zhang et al., 2012).

Even less obvious is how any of the amino acid changes in proteins influence pathways concerning upstream or downstream events, and regulation of potential positive or negative feedback loops. The proteins harboring damaging missSNP might have modulated translation. If they are still functional with the changed conformation, and continue to interact with other molecules, they must disturb dynamics of pathways or networks they participate, or complexes they form. That means what is evaluated as a damaging missSNP effect is damaging for a protein function with certainty on one level, but other levels might be also affected, yet "invisible" for current evaluation capability. For that reason, the non-damaging effects of amino acid switches have to be taken with caution, especially for proteins with no known function (Teng et al., 2009). Since the knowledge about functions of a good part of AID missSNP harboring proteins is currently incomplete, all the prediction results we obtained have to be taken only conditionally.

Majority of the AID missSNPs are situated within much conserved regions of DNA, indicating that they are structures inherited and preserved for often-unknown reasons, and not eliminated as mutations from the population. The difference between missSNP in the conserved and non-conserved regions of DNA is not fully understood either, although some evolutionary explanations might be foreseen (Stefl et al., 2013).

The strongest obstacle for the evaluation of missense SNP functional effect is whether the variant is only a tag (lead) SNP or a functional causal SNP in the assigned gene, or both. Missense SNPs are by definition evaluated as functional, causal SNPs. We found that is questionable rule because if a protein has additional missSNP in the same gene, they cannot be simultaneously evaluated for a global combined functional effect: we came across such examples within NFKBIE, GSDMB and TYK2 genes (Table 3). Double missense SNPs or "double hit" may be only evaluated by PolyPhen-2 tool as two separate events. Especially questionable is evaluation of two missense SNPs in high or perfect LD that are obviously inherited together and cannot be considered independent events. We detected this situation with gene NFKBIE: rs2233434, evaluated as benign, is in perfect LD with another NFKBIE missSNP in perfect LD and 56 nucleotides apart, undetected by any GWAS, but evaluated as damaging by PolyPhen-v2. Both missSNPs are located in the conserved region, and non-damaging missSNP has more than a dozen eQTL results linked with it (while damaging missSNP has none!) as evaluated by the HaploReg. Additional problem with rs2233434 is that it is in high LD with other missSNP in conserved

region of another gene (HaploReg-v4.1 data). The case is not limited to one population, as EUR and ASN both exert the same missSNPs.

Multiple transcript variants that are encoding different isoforms represent additional complexity when analyzing coding SNPs, synonymous or nonsynonymous. No bioinformatics tool yet is available to present or predict the complexity, especially when it is well known that only very few proteins do not have splicing variants: it is recognized the splicing variants present a base of gene complexity of the human genome. To illustrate how complex the relations between coding SNPs are, here is an example we dealt with: rs27434 is the synonymous SNP of ERAP1 genes, but there is no data on its function within ERAP1, except that is located in the highly conserved region. In LD with this cd-synSNP is another missSNP rs27044 with a damaging effect as evaluated by PolyPhen v2. In addition, rs27434 is in perfect LD with yet another synonymous SNP of ERAP1 gene, not detected by GWAS. In this and similar cases, we analyzed further only ERAP1 as a gene and its participation in networks or pathways, without knowledge on how various SNP influence on ERAP1 protein conformation and quantifying interactions. However, less frequent are cases where missSNP "improve" protein function (Stefl et al., 2013).

4.2.2. Regulatory effect of AID missense SNPs

All genes harboring missSNPs except IL6R and NOD2 have shown multiple expression quantitative trait loci (eQTLs) results obtained in multiple cell lines and tissues, adding additional weight to these missense variants. Change in the eQTL loci results in modifications of level of expression of the corresponding gene transcripts, changing mRNA and protein quantity in tissues or serum (Soubrier 2013).

In addition to change in protein binding sites, multiple regulatory characteristic have been recognized for missense SNPs, such as histone marks, DNAse binding sites and motif change. They all indicate that the missense SNPs detected by AID GWAS represent polymorphisms that influence regulatory elements with functional consequences on autoimmune diseases that are not easily understood based on available data and tools at this time.

4.2.3. Functional effects of AID synonymous coding SNPs

We evaluated a few synonymous AID SNPs (Supplemental table 2), focusing only on cd-synSNPs that have a significant potential to change regulation of gene expression. Because several cd-synSNPs belong to the genes that already contain missSNPs, we did not analyze them separately, but as a part of

the same gene (example is ERAP1). We did not have any nonsense SNPs causing stop codon that would result in aborted or partial proteins usually with no function (Ward and Kellis, 2012).

There were only nine GWAS AID synonymous SNPs (cds-syn SNPs) (Supplemental table 2). We did not have any nonsense SNPs causing stop codon that would result in aborted or partial proteins usually with no function (Ward and Kellis, 2012). We were focusing only on cd-synSNPs that have a significant potential to change regulation of gene expression.

All synonymous SNPs are evaluated by RegulomeDB for regulatory influence. Few are found in the same genes that we already included in the functional SNP gene set, the genes like NOTCH4, NOD2, CARD9 and PADI4. Only four synSNPs were scored with the highest scores.

One of these cds-syn SNP (rs1142287) belongs to the coding sequence of gene SCAMP3; it is found as a risk SNP for Crohn disease. The SCAMP3 encoded protein is an integral membrane protein; it functions as a carrier to the cell surface in post-Golgi recycling pathways and trafficking in endosomal pathways. The second cds-syn SNP (rs9858542) is found in BSN gene, but its encoded protein is very specifically expressed only in the presynaptic cytoskeleton. In addition, rs9858542 is in perfect LD with two missSNPs; one missSNP belongs to BSN gene and other one belongs to MST1 gene, the gene already on the list of significantly associated AID missSNPs. No annotation exists to explain link between BSN protein and two AID (CD and UC). The third cds-syn SNP was rs495337 found in SPATA2 gene; it is evaluated as very low impact SNP by RegulomeDB. The SPATA2 gene has no known function. There is only one study with orthologs in zebrafish, which suggests SPATA2 has a function in proliferation of beta cells in pancreas (Maran et al., 2009). The fourth synonymous SNP rs2240335 belongs to gene PADI4 and does not have any other functional SNPs in high LD. This gene is a member of a gene family that encodes enzymes responsible for the conversion of arginine residues to citrulline residues, a process very important for cutrulintaion of antibodies in RA. This gene may play a role in granulocyte and macrophage development leading to inflammation and immune response. For that reasons, we consider rs2240335 an interesting result, because PADI4 was long been on the list of AID risk factor genes.

The cd-synSNP rs3810936 belongs to TNFSF15, and has low impact, if any, on regulation. However, this gene encodes a cytokine of the tumor necrosis factor (TNF) ligand family; it can activate NF-kB pathway, but is not expressed in either B or T cells. TNFSF15 does not act in any of the known TNF pathways. We could not find any data supporting a potential role for this synSNP.

4.2.4. Functional effect of non-coding AID SNPs

Similarly to the coding SNPs, the AID non-coding SNPs (ncSNPs) could also be lead SNPs in GWAS for genes in their LD, both for coding genes or non-coding genes. Alternatively, ncSNPs might represent solely changed regulatory elements that have an effect on gene expression of cis genes. Focusing exclusively on exome and SNPs solely influencing exome, such as missense coding SNPs or synonymous coding SNPs, is an extremely serious limitation in complex trait genetics (even an error). Noncoding genetic variations play essential roles in complex traits and complex diseases, roles larger than in Mendelian genetics or in somatic cancer genetics (Kiezun et al., 2012).

We retrieved and analysed the AID SNPs in untranslated regions: UTRs, both 5' and 3' ends, and nearGene regions and in intronic regions. Variants in the 5′-UTRs may influence the promoter activity of a gene, whereas the variants in the 3′-UTRs may change the mRNA degradation rate mediated by microRNAs and RNA-binding proteins (Frazer et al., 2009). We did not specifically detect any AID SNP with a location at splice junction sites. Variants located at splice junctions may alter the splicing patterns of genes (Faustino and Cooper, 2003). Because there are no precise tools that would assign functions to ncSNPs, we consider them only as a part of genes they are located in a broader sense.

SNP variants of pseudogenes are difficult to understand, especially when it is known that pseudogenes may have a huge repercussion on the immune response (Pink et al., 2011). Among our analysed AID SNPs, we have an example of AID SNPs located in the pseudogene WASF5P (Chr.6) (Table 5). WASF5P is pseudogene of the gene WASF3 (Chr.13) that belongs to the family of genes encoding Wiskott-Aldrich syndrome (WAS) proteins (NCBI Gene). Wiskott-Aldrich syndrome is a serious disease of the immune system, an immune deficiency (Massaad et al., 2013). This gene family encodes the multiprotein complex that connects kinases and actin and serves to transduce signals that involve changes in cell shape, motility or function. However, WASF5P has no known function (Kurisu and Takenawa 2009).

4.2.5. Regulatory effect of AID ncSNP associations

ncSNPs might modify expression level of the corresponding gene transcripts, changing mRNA and protein quantity in tissues or serum (Soubrier 2013). Potential for regulatory effect of all retrieved AID ncSNPs was assessed for each variant using RegulomeDB software tool that provides scores based on available data, mainly from ENCODE project. We selected ncSNPs with highest RegulomeDB scores (from 1 to 3) as the ncSNPs with the highest potential to influence gene expression. We did not study

ncSNPs with lower scores not only because they are less likely to exert their influence but also because of their sheer number (Suppl. Table 1).

More than 35% of numerous ncSNP loci had very high RegulomeDB score (Table 6), reflecting elements with strong regulatory role like eQTL. Usually ncSNPs were annotated with regulatory elements including histone marks, DNAse sites, protein binding sites and motifs that have been changed, in addition of numerous eQTL results in several cell lines and tissues. However, it is very difficult to understand what the meaning of all these associations is, except that they do exist and for a reason must be taken into consideration. Again, no publicly available software tool is able to concatenate them into biological processes or pathways, or linked them to the known complexes. They are evaluated as the regulatory elements associated with AID, but everything in between is still kept in dark (Edwards et al., 2013). ncSNPs do not provide explanation for inheritance, nor did they fit into any mechanism for AID development. Even if we accept that they represent regulatory sequences (based on ENCODE data), there are no annotations for their possible influence on gene expression through transcriptional, posttranscriptional, and posttranslational mechanisms; no immune pathways are annotated for regulatory elements so that we could use comparison between normal (wild type) and variant (changed) regulatory elements. Only few papers have addressed this issue, referring that changed regulatory motifs could have specific functional consequences. Consensus among researchers is that experimental approaches are necessary for confirming mechanistic relevance of regulatory elements, and future functional experiments will resolve problems and predictions unveiled by the AID GWAS data (Tak and Farnham 2015).

However, we found one good example how regulatory impact of ncSNPs might be connected with the nearest genes or a gene within boundaries they locate. However, ncSNPs might also influence distant genes, keeping in mind that a LD region could be discrete and not necessarily linear. Such example is UC SNP variant rs9263739 (with $p=5 \times 10^{-67}$) located in the intron of gene CCHCR1. This ncSNP it is in perfect LD with two missSNPs in the coding regions of other two genes, PSORIS1C1 and PSORIS1C2 genes. The two missSNPs are detected in several AID (psoriasis, SLE, MS etc.) as risk factors, for both EUR and ASN populations. Because the CCHCR1 gene has no known function, we made a choice to evaluate this signal as coming from two missSNPs on two genes in its perfect LD. PSORIS1C1 and PSORIS1C2 genes are among the first genes linked to psoriasis by linkage studies and evaluated in OMIM as the strongest risk factors. However, no data existed for UC.

Other AID ncSNPs located in the 3' and 5' regulatory regions, were also very often found together with intronic or missSNPs in the same gene region (Supplemental Table 1). We did not analyse them separately, although ncSNPs located in the 3' and 5' regulatory regions are considered variants with

proven influence on transcription of corresponding genes (Okada et al., 2014). We included them directly in the list of ncSNP coding gene set, and did not analyse their potential regulatory function.

4.2.6. Genes linked to ncSNPs

Not all tag ncSNPs could be linked to regulatory effects. They might simply reflect genes in their vicinity. We searched for the potential coding genes located in the high LD with ncSNPs as lead SNPs, using HaploReg software. Almost 90% of all highly scored ncSNP have several coding or non-coding genes in their corresponding high LD (r2> 0.8). We detected 39 coding genes in high LD with ncSNPs based on the current annotations (Table 6 and 8). This set of genes, named ncSNP gene set, we further used for gene-pathway prioritization analysis (Table 5).

We also probed the lead ncSNPs for non-coding genes in high LD (r^2> 0.8), which we could not be put in any pathways or networks prioritization probes. As our results show, an unexpected abundance of non–coding genes emerged, including micro RNAs (miRNA), long noncoding intergenic RNAs (lincRNA), small nuclear RNAs (snRNA), and other uncharacterized RNAs. Only one quarter of analysed loci in LD with ncSNPs did not have any ncRNA genes or annotations data did not exist at the time (Table 6 and Table 7). Often more than one RNA genes were detected in LD for each analysed ncSNP.

Non-coding genes in high LD with ncSNPs were further analysed for their function. However, none of RNA genes could be explained for their role, because all of them, with no exceptions, were very poorly annotated. Obtained RNA genes could not be analysed as a separate gene set for their functional properties either, because there are no databases or tools (yet) that incorporate RNA genes into functional pathways (except occasionally for micro RNAs).

4.2.7. Relevance of eQTL results linked to AID SNPs

Levels of gene expression are highly heritable (Morley et al., 2004), and they are defined by eQTLs. We have detected many eQTLs for the AID SNP associations (Supplemental table 3). However, it is not clear how disease risk variants influence eQTLs. It is obvious from our results that a good number of AID SNP variants might change eQTLs in their corresponding regions. The majority of identified eQTL are *cis*-acting or local eQTL (Cheung and Spielman 2009), arbitrarily defined as regulation of genes within 1 Mb, but genetic variants can also affect the expression of genes that reside further away or are on different chromosomes, distal or trans-eQTL (Westra et al., 2013). Furthermore, the target genes of eQTL associations could be coding genes or noncoding RNAs (Kumar et al., 2013). We have found all classes of genes or regulatory elements in the vicinity of the analyzed GWAS AID SNPs. There are no software tools for searching distant genes.

4.2.8. MicroRNA genes

Among the detected ncRNA genes connected with the AID ncSNPs, we specifically analysed miRNAs because of their well-known role in gene regulation (Ventriglia et al., 2015). Several computational tools could provide information about miRNA targets. Micro RNAs are quite conserved structures that typically regulate gene expression through binding to 3′ UTRs of targeted mRNAs to direct their posttranscriptional repression (Bartel 2009). Predicting potential target genes is the major challenge in exploring miRNA function, given that a single miRNA can potentially regulate hundreds of different genes. However, using TargetScan computational tool for target prediction, we have found only a few targets for the detected miRNAs. We detected 12 miRNAs in high LD to ncSNPs. Majority of these miRNAs have known targets as found employing TargetScan (Table 7a and 7b). However, we did not find among them any miRNAs to target GWAS AID SNP coding genes, nor we could confirm that they regulate any risk factors for AID, leaving a space for further exploration of additional genes linked to AID that are influenced by the detected miRNAs.

We also searched whether any of miRNA are influenced by SNPs in 3'-UTR of genes known to be linked to AID. We were not able to find any miRNA in the associations we have selected based on used p values (Table 7c). Evidence exists that miRNAs modulate immune cell function and that level of miRNAs changes in immune cells in T1D, as that the dysregulation in immune cells can lead to immune pathology at least in T1D; however, that knowledge is still not integrated with pathways or other networks (Ventriglia G et al., 2015). The future innovative approaches are necessary for confirming not only miRNAs mechanistic relevance in AID, but also the role of so many other RNA genes of various types connected with AID SNP associations that we detected.

4.2.9. SNP function analyses and human orthologs

Comparison of orthologs to human SNPs and genes have been often used for clarification or to gain new knowledge about diseases/traits and SNPs. Although model organisms are used frequently for human disease studies, the phenotypic relevance of model organism genes that are orthologous to human disease genes remains unclear and questionable (Lehner 2013). It is impossible to accurately predict phenotypic variation based on genetic variants in orthologs. Mutations in an orthologous pair of genes do not always exhibit similar phenotypes in different species. Tested SNPs in mice were not able to produce the same disease that occurs in humans. Obviously, using orthologs in model organisms could not unequivocally prove that a SNP is the direct cause of any given association. Using orthologs, even when they resemble greatly the human counterparts (as they do in a sense of a function), their

regulation, pathways and expression patterns may differ significantly. Most of the newly selected gene targets and drug targets developed from animal models have failed in humans (Wang et al., 2015).

4.3. Network and pathway analyses

In order to find a set of genes underlying the AID phenotypes, annotations for AID gene sets have to be clustered and compared to identify groups of genes that act similarly and thus may be a part of the same pathway or network. This is achieved by gene-network and gene-pathway prioritization analyses.

We defined separately the missSNP and the ncSNP sets for easier prioritization (Table 8). We could not use less than 10 members in any group of genes for prioritization, because ten or more genes are needed for the computational tools we have used. For that reason, we could not take only a subset of a few genes for gene-pathway prioritization, even when we had an argument for it.

4.3.1. Network analyses or gene-network prioritization

Using Cytoscape software tools, we were able to find a network that consisted of eight missSNP harboring proteins and TNF (nodes) connected by edges representing any type of relations. This was a positive result because it indicated existence of biological relations between TNF and NFKBIE, TYK2, TNFAIP3, NOD2, CARD9, MICA, SH2B3 and IL23R proteins. These links have not being registered by any other used method (Figure 5), nor other SNP-harboring gene nodes tested in a same way showed any edges between themselves, indicating lack of relations. The inferred functional connections between eight missSNP genes and TNF nodes are based on integration of huge number of data gathered from various sources and databases, all of which the Cytoscape tools compiles for network analyses. All edges between the nodes are of inferred nature, deduced from text mining data and structural similarities between proteins with experimentally detected interactions (some in model organisms). However, no experimental data or publications have been indicated, so we could not refer to any publications or experimental data. That is exactly the value of the network analyses by Cytoscape: its capability to make imputations based on broad data sources. The Cytoscape visualization of the networks and their interactions is suggestive of a space of the interactive communications among the SNP harboring proteins, the communications that might be important in AID pathology.

Another significant result obtained from the analyses with Cytoscape tools was the indications that all networks constructed for the missSNP proteins and TNF are integrated into one single complex conglomerate of networks, confirming that the interactions do exist between them, directly across the

first interactants they share. If they were not a part of one unifying network, we would obtain two or more unconnected independent network structures using the same tools (Figure 4).

STRING software when used to perform network analyses, was not able to confirm the Cytoscape results (Figures 7 a, b and c). STRING constructed networks for missSNP dataset, ncSNP dataset and allSNP dataset (a third dataset compiling both sets of SNPs together). All three sets are highly enriched for PPI (Table 10), confirming that they are a coherent gene/protein set with potential common biological function(s). However, in the STRING networks, we did not find edges between TNF and missSNP genes/proteins; STRING detected edges between TNF and CD40, TRAF1 and IL10, and constructed a much smaller, partial network between TNF and these three ncSNP genes/proteins with edges that represent mainly textmining, similarity or co-expression links. The STRING networks could not be parsed in the way Cytoscape allows network parsing, so intersections or unions could not be studied by STRING.

Regarding the widely accepted theory that essential genes have more PPI partners than nonessential genes (Jeong et al., 2001), it is important to emphasize that TNF has the largest network with over 1500 genes/proteins, all first interactants (as detected by Cytoscape). This huge number of direct interactants might qualify TNF as an essential hub protein. Among the disease genes detected by GWAS (disease genes are the genes in SNP gene sets), almost all missSNP genes have significant number of interactants (10-100), all except LACC1 and RTNK2 genes that have none (Table 9). Few missSNP genes have less PPI than others (less than 50 edges), but still enough not to consider them low impact genes. However, they are not hub genes either, based on the number of interactions they can make and their expression patterns (Goh KI et al., 2007). From the expression pattern of AID missSNP genes (Figure 6), it was clear they are present in immune cells, but also in fetal tissues and several adult tissues, so they are not limited to only one type of cells or tissues (hub genes are by definition expressed in many tissues most of the time). Expression patterns of the missSNP genes are important, because if the proteins are not expressed together in time and space, they cannot obviously interact directly, assemble pathways or cooperate in any type of networks.

Within the human disease network (HDN), immunological diseases are enriched with nonessential disease genes (Goh et al., 2007; Seong KH et al., 2015). For example, PTPN22, FCGR2A and NOD2 genes are considered disease nonessential genes, although we found a huge number of interaction they make in networks. Some other highly wired genes like TNFAIP3, TYK2, SH2B3 and CARD9 are not included in the HDN list of classified genes. For the gene classification of the HDN, the major criterion is disease mortally, and the AID we studied, although have high mortality unless treated, are not

considered as such (Crohn disease, psoriasis, SLE, etc.). However, T1D is on the list of diseases with essential disease genes.

The major downside of the HDN theory is that it only takes into account OMIM based diseases and their genes. The majority of the GWAS AID genes are not found in OMIM, because the OMIM data cover only genes that were discovered in pedigree studies based on linkage analyses of monogenetic diseases, and do not cover diseases with unknown or complex genetic background. OMIM contains 2430 non-coding SNPs (0.0001% of all human SNPs) and 5327 coding genes (0.01% of all – 100-fold enrichment) out of known ~432 thousand coding SNPs in the current built of human dbSNP database (OMIM McKusick; Bromberg 2013). For that reason, the HDN contains insufficient number of disease genes and diseases. Interestingly enough, the classification of cancer genes in HDN is dubious (to say the least), because they are considered hub and essential genes, but they are still very locally expressed. In addition, there are many erroneous conclusions such as that TNFRSF1A gene is considered an essential gene for familial periodic fever disease (which is not a deadly disease and is classified into Immune system diseases). On the other hand, TNFRSF1A gene participates in human pathways majority of which we found to be crucial for AID, yet none of the missSNP AID genes is considered essential (Seong et al., 2015).

Essential human genes are likely to encode hub proteins and are expressed widely in most tissues (Goh et al., 2007). This suggests that the vast majority of disease genes are nonessential, as they show no tendency to encode hub proteins. Their expression pattern (mainly in specific tissues), indicates that they are localized in the functional periphery of the network. However, recently, when the analysis was expanded to examine the PPIs of human disease genes, it was found that essential disease genes have more PPI partners than non-essential disease genes (Seong et al., 2015). Comparison of the AID missSNP genes with results of this study is leaving unclear their status. In our opinion, AID missSNP might not be essential genes, but they are definitively high-wired genes and collectively, and their variants might be able to influence many biological processes. The examples are missSNP genes TYK2, NFKBIE and several others. Both NFKBIE and TYK2 are extremely important key enzymes in two non-overlapping essential signaling pathways, NF-kB and Jak-STAT signaling pathways that are operational in many cells and tissues, and not limited to lymphoid cells. Major role of NFKBIE gene encoded protein is inhibiting NF-kappa-B by complexing with it and trapping it in the cytoplasm. The TYK2 protein associates with the cytoplasmic domain of type I and type II cytokine receptors and promulgate cytokine signals by phosphorylating receptor subunits. It is also associated with both the type I and type III interferon signaling pathways (NCBI Gene database).

4.3.2. Pathway analyses or gene-pathway prioritization

It is rarely the case that one gene is responsible for one function, contrary to the classical view. Rather, an assembly of genes constitutes a functional module or a molecular pathway. By definition, a molecular pathway leads to some specific end-point in cellular functionality via a series of interactions between molecules in the cell. Furthermore, it is an oversimplification to view a single pathway as a discrete and independent entity. Pathways act on defined sites overlapping with other pathways compiled in modules.

In order to connect SNP data and disease phenotypes, it is essential to sort SNPs influenced genes into functional pathways and then link them to diseases (Bromberg 2013). Biologically this process makes sense, because if diseases result from pathways' breakdown, then disabling any of the pathway components can produce similar phenotypes, and genes responsible for similar diseases often participate in the same interaction networks or pathways. Some newer proposals of disease classification go so far to propose that a disease soon will be considered a collection of mischief pathways. Accordingly, the drugs (probably a cocktail of drugs) will be targeting damaged, altered pathways, and not only certain gene product, usually an enzyme. It would be important to distinguish between neutral SNP changes from pathogenic SNP changes not only on the protein level, but also on pathway level. Not every SNP, even missense SNP will change a protein towards malfunctioning. However, even less is known about what changes are needed to for a pathway to accumulate dysfunctional members, in order to go over its "tipping point", and become a disease pathway. No tool is currently available to provide this type of information yet.

We have used gene-gene (protein-protein) interaction and pathway information to prioritize candidate genes and find pathways in which AID SNPs function, but we are not able to distinguish whether such pathway is damaged or not by the presence of such variant gene/protein.

The Cytoscape encouraging results prompted us to search for connections between missSNP harboring proteins and canonical pathways. We conducted molecular pathway enrichment analysis using KEGG databases as a main pathway information resource, but we also compared the KEGG pathways with pathways from other sources, like Regulome, INC, Wikipathways, etc. of it was necessary. The goal of pathway-level methods is to determine if the gene sets, found by genetic associations from a GWAS, are enriched for any canonical pathway (gene-pathway prioritization). Extrapolation of the results would be then that a particular pathway might be a disease-causing pathway.

Majority of the AID missSNP genes participate at least in one KEGG canonical pathway (Table 13). The same ratio was detected for ncSNP gene set when manually searched for pathways in KEGG database

(Table 14). Significant number of KEGG pathways overlap between two sets of genes. Interestingly, when the SNP gene KEGG pathways were analysed for a type of pathway, two distinctive groups of pathways have emerged, signaling pathways and diseases pathways. We consider only signaling pathways the core pathways; disease pathways are rather collections of core pathways. Some pathways are collections of interactions with ligands and receptors (like networks), and illustrate a disease or network of the cytokines and the cytokine receptors. They are not pathways where the members are interacting in an directional way up to the end-point resulting molecules. We considered only core pathways that are harboring SNP genes for further analyses.

The SNP gene core pathways or signaling pathways were interrogated against TNF signaling pathways in order to find overlapping pathways. The TNF signaling pathways have been obtained also from KEGG pathway database. We consider them the biological vehicles for allocation of aTNF biologics actions, because aTNF biologics inhibit TNF wherever it acts.

After manual comparison of all SNP gene pathways and TNF pathways (Supplemental table 4), a much smaller set of nine intersecting core pathways was obtained (Supplemental table 5, Table 17). We consider these selected SNP gene/TNF pathways as a subset of pathways responsible for disease development (pathological pathways), but at the same time these pathways are also responsible for the propagation of aTNFs actions. The set of pathways is responsible for remission of AID, because they are a playfield where aTNF drugs interfere with autoimmune/inflammatory disease pathways, and regulated them down to remission status by reverting them into normal functioning pathways. We consider these pathways curative pathways. Consequently, other AID pathways, except the curative pathways, cannot be regulated by the aTNF biologics.

However, the selected pathways are not all equally influenced by TNF. For some, TNF is on a proximal side as it acts as a signal or trigger molecule and for others, TNF is on a distal side, being triggered by the pathway or is synthesized as an end-product of the pathway activation (Table 16). Toll-like receptor signaling pathway is not regulated by TNF, as TNF is not a signaling molecule in the pathway; TNF is synthesized as a consequence of activation of the toll receptors with their ligands. Toll-like receptor signaling pathway is also influenced with Jak-STAT system, the system that TNF have no common genes with it (Table 17).

Similarly, the NOD-like receptor signaling pathway is not triggered or controlled by TNF, but it does not have the Jak-STAT complex either, unlike the Toll-like receptor signaling pathway.

In the Adipocytokine signaling pathway, TNF is a major signaling molecule. However, this is a pathway connected only with the endocrine system and it is not an immune system pathway. It explains why the proinflammatory cytokine TNF has been implicated as a link between obesity and insulin resistance. It was not confirmed that Adipocytokine signaling pathway has any role in AID under study.

TNF is also very vaguely connected with Fc epsilon RI signaling pathway; it is not a signaling or control molecule, and it is only proposed to be synthesized after activation of this pathway. The member of this pathway, gene FCGR2A, is harboring missSNP detected by AID GWAS.

TNF is a product of T cell signaling pathway, but it does not influence B cell signaling pathway.

TNF is a signal of NF-kB pathway but is also augmented by it and serve as a positive feedback loop for NF-kB pathway. TNF does not influence Ubiquitin mediated proteolysis, the pathways that is directly connected by NFKBIE and NF-kB signaling pathway. Ubiquitin mediated proteolysis, a very important pathway for regulation of protein degradation, as it is directly linked with NFKBIE and its complex that undergoes ubiquitination as a regulation process. Any SNP variant in NFKBIE has a potential to change its binding to complex further degraded by Ubiquitin mediated proteolysis. Ubiquitin mediated proteolysis also has as a member the GWAS AID SNP linked gene/protein UBE2L3, so it might be one of the pathogenic AID pathways because it contains a risk factor for AID and it degrades the NFKBIE complex and release a transcription factor that travels into a nucleus.

Another interesting pathway/process is Apoptosis pathway, which is triggered by TNF, and might be considered in this group of the curative pathways, but none of AID SNPs is connected with Apoptosis. We are suggesting that at least NFKBIE gene should be included, because it is involved in the NF-kB pathway, but it is not yet annotated as such (and for that reason, we were not able to retrieve Apoptosis as a pathway for NFKBIE from KEGG, and other tools based on KEGG like Cytoscape). However, it is well accepted that apoptosis is regulated, more precisely inhibited, by the NF-kB pathway and the apoptotic process is stopped and reverted into a survival mode when NF-kB pathway is activated in a cell. In addition, both counterparts of NFKBIE, NFKBIA and NFKBIB are members of NF-kB pathway and Apoptosis. It would be interesting to see in future how missSNP in NFKBIE participate in the process: whether missSNP augments the apoptosis or it enhances inhibition of the apoptosis. Jak-STAT signaling pathway has no role in apoptosis.

As a result of the orthogonal evidence, we found that at least two sets of pathogenic AID pathways exist. Only one set of pathways is controlled by TNF and only that set of pathways can be modified by aTNF biologics (Figure 6). The second set of pathogenic pathways includes Jak-STAT signaling pathway and

Jak-STAT complex dependent signaling pathways such as the cytokines' pathways that solely employ Jak-STAT complex to propagate signals from various cytokines (TYK2 belong to JAK family of tyrosine protein kinases).

NOD-like signaling pathway has TNF as its product that augments further inflammation, together with IL6 cytokine.

Other pathways that harbor missSNP genes do not have any overlapping with two sets of pathways and act independently of TNF and Jak-STAT signaling (for example Notch signaling pathway). Another example is Regulation of autophagy, a potential GWAS AID SNP pathway with no connection with TNF.

The finding of at least two independent sets of pathways explains unresponsiveness to aTNF drugs. The pathways that are not influenced by TNF cannot be controlled by its inhibitors. It also explains how and why some of AID are better controlled, (but not cured or reversed) with a relatively new class of drugs, Xeljanz (Tofacitinib citrate). XELJANZ is a prescription drug approved by the FDA in 2012, and acts as a Janus kinase (JAK) inhibitor.

Some AID SNP pathways are influenced by both TNF and Jak-STAT complex: Osteoclast differentiation pathway is influenced by TNF as a signaling ligand, but is controlled by Jak-STAT signaling partially, dough only partially.

Disease pathways do not belong to the core signaling pathways, but their contributions to AID disease development are undisputable. They are pathological pathways, but are not controlled by aTNF drugs. Their existence is an additional burden for the AID therapy and potential cure.

We also prioritized gene-pathway using STRING (Table 18) and ConsensusPathDB (Table 19 and 20).

STRING enrichment in pathways for the missSNP gene set revealed only three signaling pathways, all three included in the set of pathways we discovered manually searching KEGG database. When the allSNP gene set was analysed, it showed additional two pathways, one of them not discovered before (Chemokine signaling pathway) in addition to already detected T cell receptor signaling pathway. We selected the STRING pathways after correction for FDR (<0.05).

When we used ConsensusPathDB for pathway enrichment analyses, 76 pathways were retrieved with 17 missSNP genes (out of 23) present at least in one pathway. Many of these pathways, coming from several databases in addition to KEGG, such as Reactome, Wikipathways, Signalink, PID, BioCarta, etc. are repetition of the same KEGG pathways. However, a few new pathways were retrieved as well: NF-

kB signaling, TNF signaling, Toll-like receptors signaling and Cytokine-cytokine receptor signaling, which are contained completely in the same KEGG collection of pathways.

Pathway enrichment by ConsensusPathDB for the ncSNP gene set has not detected any new pathways that were not previously detected for the missSNP genes set. It is interesting that similar portion of ncSNP genes (70%) participate in various pathways. Pathway enrichment for the all SNP genes did not improved with the number of genes, leaving the same portion of the SNP genes detected to participate in pathways (70%). We found that ConsensusPathDB pathway enrichment analyses provided valuable results and was easy to use. We did analyses separately enrichment for complex formation of SNP genes (new ConsensusPathDB tool), but the results for complex formation enrichment were inconsistent and difficult for interpretation, and for that reason we did not include them in our research.

The ConsensusPathDB results for AID SNP gene-pathway prioritization indicates that close to 1100 genes are engaged in the pathways in which AID GWAS SNP may operate. A bit over 70% of AID SNPs have influence on nine KEGG pathways. The rest of 30% of genes, which are harboring AID SNPs or are influenced by AID SNPs, cannot be assigned to any pathways, which still make a big portion of unknown pathways. It would be hard completely to understand the genomic variation that is represented by GWAS SNPs. For that reason, better annotation of all components must be accomplished before making any scientific conclusions about influence of the genomic variation on the phenotypic expression of autoimmune/inflammatory diseases.

However, identical results obtained using several different approaches are good indication that what is shown by our study is a true reflection of the relationship between variants and AID.

The SNP-gene-pathway prioritization analyses showed that the genes and other structures harboring mutations that qualify as AID risk factors also play pivotal roles in the pathogen defense and other life preserving immune functions (e.g. neoplastic surveyance). It might be a reason why common risk loci are so numerous and are not purified from human genomes. They are intertwining with other genes responsible for controlling infections or neoplastic formation.

4.4. Gene Ontology helps clarify the function of GWAS AID SNPs

In order to uncover the potential biological rationale between AID SNPs and disease mechanisms, we applied GO term enrichment/ or overrepresentation analyses. GO term enrichment analyses are

independent from all other used approaches. GO terms are differently structured, but the source of information is based on the same scientific facts.

Enrichments of the GO terms for the biological process (BP), molecular function (MF) and cellular components (CC) of SNP gene sets have been performed by STRING (Table 21) and ConsensusPathDB tools (Tables 22, 23 and 24). We used both Bonferroni correction for multiple testing and FDR for comparison, as the Bonferroni correction is considered too strict for GO term enrichments.

The GO enrichments have provided information on functional activity and its location of AID SNP gene sets. They indicate that AID SNP gene sets are functioning in biological processes mainly connected with the functions of immune system and their regulation. Interactions are realized through the protein-protein domains' interactions and are located on the various cell membranes, including cell surface membrane.

It is interesting that ncSNP gene set is enriched with MF term for sequence-specific DNA binding transcription factor activity, which is not found among missSNP gene set. All GO terms are consistent with previously obtained results, because no new or unforeseen process has been discovered.

More than 95% of missSNP gene set and 87% ncSNP gene sets have been classified or recognized by GO terms, which represent a significant result, in comparison to 70% of gene classification by gene or pathways prioritization methods, or even less when the AID SNPs are manually searched.

There is a huge disbalance between numbers of BP GO terms compared with MF and CC terms. It seems that it is more a systematic problem of GO terms than a bias towards BO terms caused by our selection of genes.

4.4.1. GO disease terms enrichment for AID GWAS SNP gene sets

At the end of our research, we wanted to validate the obtained results and we chose to test what enrichment in diseases, if any, could be obtained by enrichment of GO terms for diseases (Table 26). We used STRING to find the disease GO terms and were positively surprised when all SNP gene set returned enrichment in all autoimmune/inflammatory diseases we started with. Other less specific GO disease terms also have been enriched for, all in the higher category of disease classification.

In the next step, we wanted to test whether we would obtain to find the same pathways for the AID, using relatively new software tool DiseaseConnectDB, which connects diseases with pathways.

The retrieved pathways have shown the same pattern as obtained by using other tools (Table 25). We organized them using the same method for pathway selection as before (core vs. complex disease pathways) and obtained the pathway set enriched in AID. The set is completely overlapping with the pathway set results presented in the Table 17.

When the pathways found by DiseaseConnectDB tools are compared with TNF pathways (Table 17), the result shows that all signaling pathways, except Jak-STAT pathway, are intercalated with TNF action either on a receiving end or a dispatching end. It confirms the previous finding on shared pathways using other software tools in our study.

Independent of tools or databases used for analytical purposes, the results have shown that the AID GWAS SNPs participate in at least two groups of pathways and the number of core pathways is relatively small. Although we were not able to pick up any particular gene responsible for variable response/ no response of AID patient population to the anti-TNF biologics, we were able to find a very few pathways with not more than 100 genes that carry response to anti-TNF drugs.

4.5. Pleiotropy at the gene and disease level

We have the several autoimmune /inflammatory diseases under study, linked for the same missense SNPs, obviously targeting the same genes. In addition, other immune system related disease (and not only the AID under study) also demonstrated the same missense SNPs or variants in the same genes (Table 1 and 2).

We detected more than dozen shared SNP associations detected by GWAS for AID (Table 2). As the SNP-gene prioritization analyses show, many shared genes or shared risk factors followed the basic commonality among SNPs. The same phenomenon continued later when we analyzed the pathways detected for the SNP harboring or influenced genes. This was not surprising finding, because the very idea that different AID can be and are treated with the same therapy has in its base realization that the same pathways are operative in all these AID and for that reason the aTNF biologics act as successful therapy in these diseases. Common SNPs seems to be confirmation of what was already pragmatically discovered: AID diseases share pathological pathways, the pathological pathways are similarly distorted by their protein members that are as proteins damaged by the mutations in their coding genes. What is new is that the coding variants are consequence of common SNP variants that have been detected by GWAS. The variants are not just rare mutations, but most probably variant “load” that exist in human populations, which under certain circumstances become active risk factors. The phenomenon is called

pleiotropy, and it obviously exists at the several levels: SNP variants, genes, pathways, pathway or network modules and diseases.

Allelic pleiotropy, where one genetic variant influences several distinct phenotypes, is increasingly recognized as a common phenomenon from GWAS findings, especially for immune-mediated diseases (Solovieff et al., 2013).

Furthermore, it was found that the same gene, with the different SNPs, is associated with the multiple diseases. However, it was impossible to evaluate whether the other SNPs contribute or not to the same functional effect as the missense SNPs on the same gene, even when they are missSNPs as well. We have similar case for the gene encoding IL7R (Table 2). Allelic pleiotropy is not only linked to missense SNPs, as non-coding SNPs are also shared among common diseases.

There is no bioinformatics tools currently that can compare influence of the different GWAS variants on the function of the same protein, gene, or DNA regulatory structure. It would be very important issue, as the existing data show completely different influence of the same variants in GWAS studies: the same SNP is protective in one disease and detrimental in another disease (Cho and Gregersen 2011; Wang et al., 2015). No explanation is provided, but it might be connected with other participants of the pathways or networks they are engaged in.

Investigation of pleiotropic effects could better inform on disease biology and predict potential adverse events of derived targets (Hebbring et al., 2014). The leading signals of association might be the exact same tag SNPs, but whether these GWAS association signals in AID refer to the same or different distinct causal variants or causal genes, remains unclear (Diogo et al., 2014).

The network analyses research has found that if a gene has more connections in the cellular network (not necessarily only PPI, but also co-expression, co-localization, co-regulation interactions) or if the gene is a hub gene, then its perturbation tends to result in the disconnection of multiple cellular functions, which leads to disease pleiotropy (Bromberg 2013).

An additional approach to investigate pleiotropy comprehensively is through genotype data linked to clinical data derived from electronic medical records. This unbiased approach, called phenome-wide association study (PheWAS), allows for genotypes of interest to be tested for association to hundreds of clinically relevant phenotypes (Hebbring et al., 2014).

The pleiotropy of SNP-gene and pathways is obvious at the disease symptoms level, so that some current studies even suggest a reclassification of diseases according to the pathways, which underlie a

disease. The diseases might be considered as perturbations of pathways or pathway/network modules that have reflected in signs and symptoms of the diseases. This new approach might also have a positive effect on therapy and discovery of new treatments because new drugs would be correcting not only function of one gene or its product (for example an enzyme), but they would be able to modulate a whole pathway, reestablishing its normal activity (Costapas and Hafler 2013).

5. CONCLUSION

In our research, no individual single gene was discovered to be responsible for therapeutic effect of the aTNFs, as it might be expected if compared with some reports on other (simple) drugs. Instead, the four major canonical signaling pathways have been detected as the most probable carriers of the response to aTNFs, making the difference between achieving remission in some AID patients and lack of therapeutic response in others. They are NF-kB signaling pathway, Antigen processing and presentation pathway, T cell receptor signaling pathway and TNF signaling pathway. These four pathways are structured and affected by the hundreds of genes. The genes influenced by the risk GWAS SNP variants associated with the AID phenotypes are among these genes. The variants of these genes/proteins may modulate these pathways in many yet unknown ways. Based on the common knowledge that proteins overwhelmingly are functionally damaged by missSNP, the AID missSNPs most probably can only provoke disregulation and disturb the four pathways' flow negatively, conditioning disease development on one side and permitting intervention by aTNFs on the other side. All other non-missense risk SNPs also carry potential to contribute to the processes, but at the level that cannot be evaluated at this time.

Our results indicate that the AID pathways, ones that are responsible for pathogenesis of the diseases, belong to at least two groups of pathways. Two separate groups of pathogenic disease pathways may be redundant in the pathogenesis of certain AID. One group of pathway(s) may be more active in a subset of individual AID patients based on their particular risk SNPs, while the second group of pathway(s) may be active in others, thus explaining the lack of aTNF clinical response in that subgroup of AID patients. Balance between pathogenic pathways in each subject dictates the response to aTNF.

The detected regulatory elements and non-coding genes that we found to be relevant for AID pathogenesis, could not be fitted into any pathways, because there are no bioinformatics tools to incorporate regulatory elements and non-coding gene products into existing canonical biological processes. Finding ways to incorporate them into functional biological processes represents a future next step in understanding of their biology.

By comprehending the study problem, finding ways to answer it and solving methodological problems along the way, we gained the comprehensive methodological knowledge how to reveal the biological sense behind the GWAS AID SNPs. This comprehensive knowledge could be beneficial for analyses of complex pathogenesis of other common diseases as well as potential intervention pathways that should be targeted by drugs (new drugs or existing repurposed drugs). As the result of our research, new

problems are detected, opening avenues worth further investigation, and leaving space for deepening our understanding of the complex diseases such as AID.

Gene Ontology emerged as a very helpful approach and an easy to use organization structure that might directly provide enrichment of descriptive data for genes of interest. All GO aspects are useful for analyzing underling biology of genes with GWAS detected variability in human genomes. However, GO is currently biased towards rich domain of biological processes, leaving poorly described molecular functions and especially cellular compartmentalization, and even less integrated pathways or diseases descriptors. Other similar ontologies of non-gene genetic elements are needed to help solve complex questions like one we dealt with in our research.

Based on our experience in application of bioinformatics software tools, we concluded that the existing bioinformatics tools and databases might provide only partial answers. Mutual obstacles are lack of complete annotations in the current public databases and insufficient integration of tools, as they are still not matching the complexity of the biological problems.

Our research highlighted the state of the current annotations in public bioinformatics databases, which need much more information and refinement in order to produce relevant conclusions for clinical problems similar the one we investigated. All databases suffer from inconsistency and incompleteness. It is almost disappointing how huge the portion of genes is with no or very scarce annotations. Little is known about non-coding genes, and almost nothing about pseudogenes; participation of non-coding genes in the biological processes is almost completely a void space.

There is also a great need for new, improved and enriched bioinformatics software tools; especially tools that are able to explore simultaneous influence of SNPs and genes, as well as to analyse potential polymorphism of pathways directly based on SNP polymorphism. It is also obvious that many tools suffer from oversimplification. Existing software tools, although exceptionally operative for the purpose they are invented, have very limited capacity for complex research needed in biology and medicine.

6. REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–1073.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S,McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-249 (2010).

Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Current protocols in human genetics / editorial board, Jonathan L Haines et al., 2013; 07:Unit7.20. doi:10.1002/0471142905.hg0720s76.

Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;4. doi: 10.7554/eLife.05005.

Agarwal P, Owzar K. Next Generation Distributed Computing for Cancer Research. Cancer Informatics. 2014;13(Suppl 7):97-109

Aggarwal BB, Gupta SC, Kim JH. Historical perspectives on tumor necrosis factor and its superfamily: 25 years later, a golden journey. Blood. 2012;119(3):651-665.

Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science.2008;322 (5903): 881-8.

Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Research 2009; 37:D793–796.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet.2000;25:25–29.

Banchereau R, Cepika A-M, Pascual V. Systems Approaches to Human Autoimmune Diseases. Current opinion in immunology. 2013;25(5):598-605.

Barabási AL, Albert R. Emergence of Scaling in Random Networks. Science.1999;286:509–512.

Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56-68.

Barshir R, Shwartz O, Smoly IY Yeger-Lotem E. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. PLoS Comput Biol. 2014;10(6):e1003632.

Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215–233.

Beaudoin M, Goyette P, Boucher G, Lo KS, Rivas MA, Stevens C, et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. PLoS Genet. 2013;9: e1003723.

Behm BW and Bickston SJ.Tumor necrosis factor-alpha antibody for maintenance of remission in Crohn's disease. Cochrane Database Syst Rev. 2008;(1):CD006893.

Ben-Horin S, Kopylov U, Chowers Y. Optimizing anti-TNF treatments in inflammatory bowel disease. Autoimmun Rev. 2014;13(1):24-30.

Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc (B) 1995;57:289–300.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature.2008;456:53–9.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009; 25(8):1091-3.

Blake JA. Ten Tips for Using the Gene Ontology. PLoS Comput Biol 2013;9(11): e1003343.

Bogdanos DP, Smyk DS, Rigopoulou EI, Mytilinaiou MG, Heneghan MA, Selmi C, Gershwin ME. Twin studies in autoimmune disease:genetics, gender and environment. J Autoimmun.2012;38(2-3):J156-69.

Borecki IB, Province MA. Genetic and genomic discovery using family studies. Circulation. 2008;118 (10):1057-63.

Bossi A, Lehner B. Tissue specificity and the human protein interaction network. Mol Syst Biol. 2009; 5:260.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790-7.

Brinkworth JF and Barreiro LB. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. Curr Opin Immunol. 2014;31:66-78.

British Thoracic Society Standards of Care Committee. BTS recommendations for assessing risk and for managing Mycobacterium tuberculosis infection and disease in patients due to start anti-TNF-alpha treatment. Thorax. 2005;60(10):800-5.

Brodin P, Jojic V, Gao T, et al., Variation in the human immune system is largely driven by non-heritable influences. Cell. 2015;160(0):37-47. doi:10.1016/j.cell.2014.12.020.

Bromberg Y Chapter 15: Disease Gene Prioritization. PLoS Comput Biol 2013;9(4): e1002902.

Brunet TDP and Doolittle WF. Getting "function" right. Proc Natl Acad Sci USA. 2014; 111(33): E3365.

Buchanan CC, Torstenson ES, Bush WS, Ritchie MD. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inform Assoc. 2012;19(2):289-94.

Busard C, Zweegers J, Limpens J, Langendam M, Spuls PI. Combined use of systemic agents for psoriasis: a systematic review. JAMA Dermatol. 2014;150(11):1213-20.

Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 2012; 44(8): 841–847.

Cárdenas-Roldán J, Rojas-Villarraga A, Anaya JM. How do autoimmune diseases cluster in families? A systematic review and meta-analysis. BMC Med. 2013;11:73.

Carter H, Hofree M, Ideker T. Genotype to phenotype via network analysis. Curr Opin Genet Dev. 2013;23(6):611-21.

Cavallari LH, Shin J, Perera MA. Role of pharmacogenomics in the management of traditional and novel oral anticoagulants. Pharmacotherapy. 2011;31:1192-207.

Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, Billing-Ross P, Madar A, Spritz R, Keinan A. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PLoSOne. 2014;9(12):e113684. eCollection 2014.

Chen GK, Jorgenson E, and Witte JS. An empirical evaluation of the common disease-common variant hypothesis. BMC Proc. 2007; 1(Suppl 1): S5.

Cheung V.G., Spielman R.S. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat. Rev. Genet. 2009;10:595–604.

Cho JH, Gregersen PK. Genomics and the Multifactorial Nature of Human Autoimmune Disease. N Engl J Med 2011; 365:1612-1623.

Cooper GS, Bynum ML, Somers EC. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. J Autoimmun. 2009;33(3-4):197-207.

Cooper GS and Stroehla BC. The epidemiology of autoimmune diseases. Autoimmunity Reviews. 2003;2 (3):119–125.

Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011;13(1):101.

Cotsapas C, Hafler DA. Immune-mediated disease genetics: the shared basis of pathogenesis. Trends Immunol. 2013 Jan;34(1):22-6.

Crisswell L.A., Pfeiffer KA, Lum, RFB, et al., Analysis of families in the Multiple Autoimmune Disease Genetics Consortium (MADGC) Collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. Am J Hum Genet. 2005;76: 561–571.

Croft M, Benedict CA, Ware CF. Clinical targeting of the TNF and TNFR superfamilies. Nat Rev Drug Discov. 2013;12(2):147-168.

Czyz W, Morahan JM, Ebers GC, Ramagopalan SV. Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. BMC Med. 2012;10:93.

Davis MM. A prescription for human immunology. Immunity. 2008;29:835–8.

Daly MJ, Carroll MC, Stevens B, McCarroll SA. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530 (7589):177-83.

Diogo D, Okada Y, Plenge RM. Genome-wide association studies to advance our understanding of critical cell types and pathways in rheumatoid arthritis: recent findings and challenges. Curr Opin Rheumatol. 2014;26(1):85-92.

Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, Del Rio NR et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics. 2014;5(1):14.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al., An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

Eberle MA, Ng PC, Kuhn K, Zhou L, Peiffer DA, Galver L, Viaud-Martinez KA, Lawley CT, Gunderson KL, Shen R, Murray SS. Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genet. 2007;3(10):1827-37.

Edwards SL, SL, Beesley J, French JD, and Dunning AM. Beyond GWASs: Illuminating the Dark Road from Association to Function Am J Hum Genet. 2013;93(5): 779–797.

Emery P. Optimizing outcomes in patients with rheumatoid arthritis and an inadequate response to anti-TNF treatment. Rheumatology (Oxford). 2012;51 Suppl 5:v22-30.

ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science.2004; 306:636–640.

ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447(7146): 799–816.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T,Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–49.

Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12:861–874.

Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res. 2008;17:347–388.

Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al., Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518:337–43.

Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev 2003;17: 419–437.

Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics. 2011;27(13):1741-8.

Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, Rioux JD. Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genet. 2008;4(4):e1000024.

Fodil N, Langlais D, Gros P. Primary Immunodeficiencies and Inflammatory Disease: A Growing Genetic Intersection. Trends Immunol. 2016;37(2):126-40.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41(Database issue):D808-15.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science 2002;296(5568):750-2.

Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009; 10: 241–251.

Fritsche LG, Fariss RN, Stambolian D, Abecasis GR, Curcio CA, Swaroop A. Age-Related Macular Degeneration: Genetics and Biology Coming Together. Annual review of genomics and human genetics. 2014;15:151-171. doi:10.1146/annurev-genom-090413-025610.

Fu J, Wolfs MG, Deelen P, Westra HJ, Fehrmann RS, Te Meerman GJ, Buurman WA, Rensen SS, Groen HJ, Weersma RK, et al., Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 2012;8:e1002431.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al., The structure of haplotype blocks in the human genome. Science. 2002;296:2225–2229.

Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol. 2008;32(4):361-9.

Garcês S, Demengeot J, Benito-Garcia E. The immunogenicity of anti-TNF therapy in immune-mediated inflammatory diseases: a systematic review of the literature with a meta-analysis. Ann Rheum Dis. 2013;72(12):1947-55.

Gene Ontology Consortium (over 160 collaborators). Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25-9.

Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43 (Database issue):D1049-56.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012; 489 (7414):91–100.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA. 2007;104(21):8685-90.

Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009;360 (17): 1696-8.

Goris A, Liston A. The immunogenetic architecture of autoimmune disease. Cold Spring Harb Perspect Biol. 2012; 1;4(3).pii: a007260.

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, Kanehisa M. Organizing and computing metabolic pathway data in terms of binary relations. Pac Symp Biocomput. 1997:175–186.

Green ED, Guyer MS, Manolio TA, Peterson JL. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. Nature.2011;470 (7333): 204-13.

Gruber TR. Ontology in The Encyclopedia of Database Systems. Ling L and Tamer Özsu (Eds.), Springer-Verlag, 2009.

Gruber TR. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 1993;5 (2):199-220.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multi tissue gene regulation in humans. Science 2015; 348: 648-660.

Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet. 2012; 13(9):654-66.

Han B, Kang HM, Eskin E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet. 2009;4:e1000456.

Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. Immunology. 2014;141: 157–165.

Hinch A.G., Tandon A., Patterson N., Song Y., Rohland N., Palmer C.D., Chen G.K., Wang K., Buxbaum S.G., Akylbekova E.L. The landscape of recombination in African Americans. Nature. 2011;476: 170–175.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009;106(23):9362-7.

Hindorff LA, Gillanders EM, Manolio TA. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis. 2011;32:945–954.

Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, et al. (2013) A Catalog of Published Genome-wide Association Studies. Available: http://www.genome.gov/gwastudies/.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al., Whole-genome patterns of common DNA variation in three human populations. Science. 2005; 307: 1072–1079.

Hirschhorn JN. Genomewide association studies: illuminating biologic pathways. N Engl J Med.2009; 360 (17):1699-701.

Hua C, Barnetche T, Combe B, Morel J. Effect of methotrexate, anti-tumor necrosis factor α, and rituximab on the immune response to influenza and pneumococcal vaccines in patients with rheumatoid: a systematic review and meta-analysis. Arthritis Care Res. 2014;66(7):1016-26.

Huang Q. Genetic study of complex diseases in the post-GWAS era. J Genet Genomics. 2015;42(3): 87-98.

Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, Barrett JC, Blackburn H, Brand O, et al., Burren O, Capon F. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature. 2013;498:232–235.

International HapMap Consortium.The International HapMapProject.Nature.2003;426(6968):789-96.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–945.

International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437 (7063): 1299-320.

International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.

International HapMap Consortium. Sabeti PC, Varilly P, Fry B, et al., Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913-918.

International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, et al., Integrating common and rare genetic variation in diverse human populations.Nature.2010;467(7311):52-58.

ENCODE Project Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). Becker PB, ed. PLoS Biology. 2011;9(4):e1001046.

ENCODE Project Consortium (plus 596 collaborators). An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012;489 (7414):57-74.

Janeway CJ Jr, Paul Travers P, Walport M and Shlomchik MJ. The Immune System in Health and Disease. Chapters 11,12 and 13. In "Immunobiology" ed. Janeway CJ Jr; 5th Edition New York: Garland Science; 2001.

Jacobson DL, Gange SJ, Rose NR, Graham NM. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol. 1997;84(3):223-43.

Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001; 411: 41–42.

Jordan DM, Ramensky VE, Sunyaev SR. Human allelic variation: perspective from protein function, structure, and evolution. Curr Opin Struct Biol. 2010;20:342–350.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33 (Database issue):D428-32.

Kamal KM, Madhavan SS, Hornsby JA, Miller LA, Kavookjian J, Scott V. Use of tumor necrosis factor inhibitors in rheumatoid arthritis: a national survey of practicing United States rheumatologists. Joint Bone Spine 2006;73:718-24.

Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: a database for integrating human functional interaction networks. Nucleic Acids Res. 2009;37: D623–D628.

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011;39 (Database issue):D712-7.

Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013;41(Database issue):D793-800.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109-14.

Kanehisa M, Goto S, Sato Y, Kawashima F, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42: D199–D205.

Kellis M, Wold B, Snyder MP, Bernstein BE, et al., (over 20 authors). Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA. 2014;111(17):6131-8.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM, Sunyaev SR. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44(6):623-630.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40:D841–D846.

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375

Knight JC. Approaches for establishing the function of regulatory genetic variants involved in disease. Genome Med. 2014; 6(10): 92-4.

Kolb WR, Granger GA. Lymphocyte in vitro cytotoxicity: characterization of human lymphotoxin. Proc Natl Acad Sci U S A. 1968 Dec; 61(4): 1250–1255.

Korta DZ, Ochieng P, Fishman D, Katz SE. Pulmonary sarcoidosis and latent tuberculosis in a patient with psoriasis treated with adalimumab. Dermatol Online J. 2015;21(1).

Kubo M, Hata J, Ninomiya T, Matsuda K, Yonemoto K, Nakano T, Matsushita T, et al. A nonsynonymous SNP in PRKCH (protein kinase C eta) increases the risk of cerebral infarction. Nat Genet 2007;39:212–217.

Kumar V., Westra H.J., Karjalainen J., Zhernakova D.V., Esko T., Hrdlickova B., Almeida R., Zhernakova A., Reinmaa E., Võsa U. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 2013;9:e1003201.

Kurisu S, Takenawa T. The WASP and WAVE family proteins.Genome Biology. 2009;10(6):226.

Lander ES. The new genomics: global views of biology. Science. 1996;274(5287):536-9.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470 (7333) :187-97.

Lander paper in New Engl J on intervention?

Lawson MM, Thomas AG, Akobeng AK. Tumour necrosis factor alpha blocking agents for induction of remission in ulcerative colitis. Cochrane Database Syst Rev. 2006; (3):CD005112.

Lazakidou, Athina. Web-Based Applications in Healthcare and Biomedicine. Ed: Lazakidou, A; Springer 2010; 7:143-156.

Lee S, Gonçalo R. Abecasis GR, Boehnke M, and Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014; 95(1): 5–23.

Lehner B. Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet. 2013;14: 168–178.

Lilley J, Wallace C. A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. PLoS Genet. 2015;11(2):e1004926. eCollection 2015.

Lin YT and Lee WC. Importance of presenting the variability of the false discovery rate control. BMC Genet. 2015;16:97.

Liu CC, Tseng YT, Li W, Wu CY, Mayzus I, Rzhetsky A, Sun F, Waterman M, ChenJ, Chaudhary PM, Loscalzo J,Crandall E, Zhou XJ. DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic Acids Res. 2014;42 (Web Server issue):W137-46.

Lucas CL, Lenardo MJ. Identifying genetic determinants of autoimmunity and immune dysregulation. Curr Opin Immunol. 2015;37:28-33.

Limdi N, Veenstra D. Warfarin Pharmacogenetics. Pharmacotherapy. 2008;28(9):1084-1097.

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

Maini RN, Brennan FM, Williams R, Chu CQ, Cope AP, Gibbons D, Elliott M, Feldmann M. TNF-alpha in rheumatoid arthritis and prospects of anti-TNF therapy. Clin Exp Rheumatol. 1993;11 Suppl 8:S173-5.

Malladi VS, Erickson DT, Podduturi NR, et al. Ontology application and use at the ENCODE DCC. Database: The Journal of Biological Databases and Curation. 2015;2015: bav010.

Malottki K, Barton P, Tsourapas A, Uthman AO, Liu Z, Routh K, et al., Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a tumour necrosis factor inhibitor: a systematic review and economic evaluation. Health Technol Assess. 2011;15 (14):271-278.

Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease.J Clin Invest. 2008 May;118(5):1590-605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al., Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753.

Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med. 2009;60:443-56.

Maran C, Tassone E, Masola V, Onisto M. The Story of SPATA2 (Spermatogenesis-Associated Protein 2): From Sertoli Cells to Pancreatic Beta-Cells. Curr Genomics. 2009; 10(5):361-3.

Marson A, Housley WJ, Hafler DA. Genetic basis of autoimmunity. J Clin Invest. 2015; 125 (6):2234-41.

Marquez A, Ferreiro-Iglesias A, Dávila-Fajardo CL, et al. Lack of validation of genetic variants associated with anti–tumor necrosis factor therapy response in rheumatoid arthritis: a genome-wide association study replication and meta-analysis. *Arthritis Research & Therapy*. 2014;16(2):R66.

Massaad MJ, Ramesh N, Geha RS. Wiskott-Aldrich syndrome: a comprehensive review. Ann N Y Acad Sci. 2013;1285:26–43.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al., Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–1195.

Maxwell LJ, Zochling J, Boonen A, Singh JA, Veras MM, Tanjong Ghogomu E, Benkhalti Jandu M, Tugwell P, Wells GA. TNF-alpha inhibitors for ankylosing spondylitis. Cochrane Database Syst Rev. 2015;4:CD005468.

McClellan J and King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210-217.

McKusick V.A. Nathans Institute of Genetic Medicine, Johns Hopkins University and National Center for Biotechnology Information, National Library of Medicine. Online Mendelian Inheritance in Man, OMIM. Available at: http://www.ncbi.nlm.nih.gov/omim

Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8(8):1551-66.

Michaud TL, Rho YH, Shamliyan T, Kuntz KM, Choi HK. The comparative safety of tumor necrosis factor inhibitors in rheumatoid arthritis: a meta-analysis update of 44 trials. Am J Med. 2014;127 (12):1208-32.

Miller FW, Cooper RG, Vencovsky J, et al., Genome-wide Association Study of Dermatomyositis Reveals Genetic Overlap with other Autoimmune Disorders. Arthritis and rheumatism. 2013;65(12):3239-3247.

Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. BioMed Research International. 2014: 13; 4023.

Monaco C, Nanchahal J, Taylor P, Feldmann M. Anti-TNF therapy: past, present and future. International Immunology. 2015;27(1):55-62.

Mooney MA, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. Trends Genet. 2014;30(9):390-400.

Morley M., Molony C.M., Weber T.M., Devlin J.L., Ewens K.G., Spielman R.S., Cheung V.G. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430: 743–747.

Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. 2015 Apr 7;11(4):e1004969.eCollection 2015.

Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B; NCBO team. The National Center for Biomedical Ontology. J Am Med Inform Assoc. 2012;19 (2):190-5.

Nanau RM, Neuman MG. Safety of anti-tumor necrosis factor therapies in arthritis patients. J Pharm Pharm Sci. 2014;17(3):324-61.

Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. Front Neuroendocrinol. 2014;35(3):347-69.

Nguyen T and Wu JJ. Relationship between tumor necrosis factor-α inhibitors and cardiovascular disease in psoriasis: a review. Perm J. 2014 ;18(1):49-54.

Novembre J., Johnson T., Bryc K., Kutalik Z., Boyko A.R., Auton A., et al. Genes mirror geography within Europe. Nature. 2008;456:98–101.

Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 37(Web Server issue):W170-3.

Nucleic Acids Research Database issue. 2015;43 (Database issue):D1-5. The 2015 Nucleic Acids Research Database Issue and molecular biology database collection.

Ober C, Loisel Da, Gilad Y. Sex-specific genetic architecture of human disease. Nat Rev Genetics 2008;9:911–922.

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506(7488):376–81.

Olivier M. A haplotype map of the human genome. Physiol Genomics. 2003;13(1):3-9.

Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. 2008;32, 381–385.

Peng K., Xu W., Zheng J., Huang K., Wang H., Tong J., Lin Z., Liu J., Cheng W., Fu D., et al. The Disease and Gene Annotations (DGA): an annotation resource for human disease. Nucleic Acids Res. 2013;41:D553–560.

Peppercorn J, Shapira I, Deshields T, et al. Ethical aspects of participation in the database of genotypes and phenotypes of the National Center for Biotechnology Information: the Cancer and Leukemia Group B Experience. Cancer. 2012;118 (20): 5060-8.

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6(7):e184.

Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Francisco Carter DR. Pseudogenes: Pseudo-functional or key regulators in health and disease? RNA. 2011;17(5):792-798.

Poliseno L, Marranci A, Pandolfi PP. Pseudogenes in Human Cancer. Front Med. 2015;2:68.eCollection 2015.

Pollard KM. Environment, autoantibodies, and autoimmunity. Front Immunol. 2015 Feb 11;6:60. doi: 10.3389/fimmu.2015.00060. eCollection 2015.

Probert L, Eugster HP, Akassoglou K, Bauer J, Frei K, Lassmann H, Fontana A. TNFR1 signalling is critical for the development of demyelination and the limitation of T-cell responses during immune-mediated CNS disease. Brain. 2000;123 (10):2005-19.

Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet. 2012;28(7):323–32.

Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. Common risk alleles for inflammatory diseases are targets of recent positive selection. Am J Hum Genet. 2013;92(4):517-29.

Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, Feolo M, Hindorff LA. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur. J. Hum. Genet. 2013;22:144–147.

Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D. MalaCards: an integrated compendium for diseases and their annotation. Database (Oxford) 2013:bat018.

Reich D, Lander ES. On the allelic spectrum of human disease. Trends Genet. 2001;17: 502-510.

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science.1996; 273:1516–7.

Rodgers M, Epstein D, Bojke L, Yang H, et al., Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis: a systematic review and economic evaluation. Health Technol Assess. 2011;15(10):i-xxi, 1-29.

Rodriguez-Fontenla C, Calaza M, Gonzalez A. Genetic distance as an alternative to physical distance for definition of gene units in association studies. BMC Genomics. 2014;15(1):408.

Ruddle NH, Waksman BH. Cytotoxicity mediated by soluble antigen and lymphocytes in delayed hypersensitivity. J Exp Med. 1968; 128(6): 1267–1279.

Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, Sim I, Chute CG, Solbrig H, Storey MA, Smith B, Day-Richter J, Noy NF, Musen MA. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS. 2006; 10 (2):185-98.

Saeed R and Deane CM. Protein-protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics 2006;7:128.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T. A travel guide to Cytoscape plugins. NatureMethods 2012; 9(11):1069–1076.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2012;40: D13–D25.

Schaid DJ, Sinnwell JP, Jenkins GD, McDonnell SK, Ingle JN, Kubo M, et al. Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. Genet Epidemiol. 2012;36(1):3-16.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22(9):1748-59.

Schett G, Elewaut D, McInnes IB, Dayer JM, Neurath MF. How cytokine networks fuel inflammation: Toward a cytokine-based disease taxonomy. Nat Med. 2013;19(7):822-4.

Schierding W, Cutfield WS, O'Sullivan JM. The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell. Front Genet. 2014;5:39.

Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. Rare Allele Hypotheses for Complex Diseases. Cur Opin Gen Dev. 2009;19(3):212-219. doi:10.1016/j.gde.2009.04.010.

Selmi C1, Lu Q, Humble MC.J. Heritability versus the role of the environment in autoimmunity. Autoimmun. 2012;39(4):249-52.

Seong KH, Kim I, Hwang J, and Kim S. Network Modules of the Cross-Species Genotype-Phenotype Map Reflect the Clinical Severity of Human Diseases. PLoS One. 2015; 10(8): e0136300.

Shah NH, Cole T, and Musen MA. Chapter 9: Analyses Using Disease Ontologies. PLoS Comput Biol. 2012; 8(12): e1002827.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13 (11):2498–2504.

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

Singh J, Furst D, Bharat A, et al. Update 2012 of the American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. Arthritis Care Res 2012;64:625-39.

Sivamani RK, Goodarzi H, Garcia MS, et al. Biologic therapies in the treatment of psoriasis: a comprehensive evidence-based basic science and clinical review and a practical guide to tuberculosis monitoring. Clin Rev Allergy Immunol. 2013;44(2):121-40.

Smilek, DE and St Clair, EW. Solving the puzzle of autoimmunity: critical questions. F1000Prime Rep. 2015;7:17. eCollection 2015.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnology 2007;25(11):1251-5.

Smolen JS, van der Heijde D, Machold KP, et al. Proposal for a new nomenclature of disease-modifying antirheumatic drugs. Ann Rheum Dis 2014;73:3–5.

Smolen JS, Landewé R, Breedveld FC, Buch M, Burmester G, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2013 update. Ann Rheum Dis. 2014; 73(3):492-509.

Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013;14: 483–495.

Sollid LM, Pos W, Wucherpfennig KW. Molecular Mechanisms for Contribution of MHC Molecules to Autoimmune Diseases. Current opinion in immunology. 2014;0:24-30.

Somers EC, Thomas SL, Smeeth L, Hall AJ. Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder? Am J.Epidemiol. 2009;169:749–55.

Song Y, Buchwald P. TNF Superfamily Protein–Protein Interactions: Feasibility of Small-Molecule Modulation. Curr Drug Targets. 2015;16(4):393-408.

Steinberg MH and Adewoye AH. Modifier genes and sickle cell anemia. Curr Opin Hematol. 2006;13 (3):131-6.

Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. J Mol Biol. 2013;425(21):3919-3936.

Storey JD and Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003;100:9440–9445.

Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. Curr Protoc Bioinformatics. 2014;47:8.13.1-8.13.24.

Soubrier F. From an ACE polymorphism to genome-wide searches for eQTL. J Clin Invest. 2013;123 (1):111-112.

Sun YV. Integration of biological networks and pathways with genetic association studies. Hum Genet. 2012;131(10):1677-86.

Sunyaev SR. Inferring causality and functional significance of human coding DNA variants. Hum Mol Genet.2012;21:R10–R17.

Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, Bassendine M, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet. 2009;41(10):1100-4.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39 (Database issue):D561-D568.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan;43 (Database issue): D447-D452.

Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin. 2015;8:57.

Taylor PC. Anti-TNFalpha therapy for rheumatoid arthritis: an update. Intern Med.2003; 42(1):15-20.

Teng S, Madej T, Panchenko A, Alexov E. Modeling Effects of Human Single Nucleotide Polymorphisms on Protein-Protein Interactions. Biophysical Journal. 2009;96(6):2178-2188.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science. 2012;337: 64–69.

Terao C, Yamada R, Ohmura K, et al. The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. Hum Mol Genet. 2011;20 (13):2680-5.

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011;43(12):1193-201.

Turner CF, Pan H, Silk GW, Ardini MA, Bakalov V, Bryant S, Cantor S, Chang KY, et al. The NIDDK Central Repository at 8 years: ambition, revision, use and impact. Database (Oxford). 2011:bar043.

Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 2003;423: 506–511.

Umićević MM, Cui J, Vermeulen SH, Stahl EA, Toonen EJ, Makkinje RR, Lee AT, et al. Genome-wide association analysis of anti-TNF drug response in patients with rheumatoid arthritis. Ann Rheum Dis. 2013 Aug;72(8):1375-81.

van der Sijde MR, Ng A, Fu J. Systems genetics: From GWAS to disease pathways. Biochim Biophys Acta. 2014;1842(10):1903-1909.

van Schouwenburg PA, Krieckaert CL, Rispens T, Aarden L, Wolbink GJ, Wouters D. Long-term measurement of anti-adalimumab using pH-shift-anti-idiotype antigen binding test shows predictive value and transient antibody formation. Ann Rheum Dis. 2013;72(10):1680-6.

Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nature Methods. 2008;6:83–90.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291:1304–1351.

Ventriglia G, Nigi L, Sebastiani G, Dotta F. MicroRNAs: Novel Players in the Dialogue between Pancreatic Islets and Immune System in Autoimmune Diabetes. Biomed Res Int. 2015;2015:749734.

Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. Cell. 2011;144(6):986-98.

Visscher PM, Brown M, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. The American Society of Human Genetics. 2012;90: 7–24.

Vyse TJ, Todd JA. Genetic analysis of autoimmune disease. Cell. 1996; 85: 311–318

Walsh SJ, Rau LM. Autoimmune diseases: a leading cause of death among young and middle-aged women in the United States. Am J Public Health. 2000;90(9):1463-6.

Wang L, Liu H, Jiao Y, Wang E, Clark SH, Postlethwaite AE, Gu W, Chen H. et al. Differences between Mice and Humans in Regulation and the Molecular Network of Collagen, Type III, Alpha-1 at the Gene Expression Level: Obstacles that Translational Research Must Overcome. Int J Mol Sci. 2015; 16 (7):15031-56.

Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Brief Funct Genomics. 2011;10:280–293.

Wang L, Wang FS, Gershwin ME. Human autoimmune diseases: a comprehensive update. J Intern Med. 2015;278(4):369-95.

Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol 2012;30:1095–1106.

Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930–D934.

Waterfield M and Anderson MS. Clues to immune tolerance: the monogenic autoimmune syndromes. Ann N Y Acad Sci. 2010;1214:138-55.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-678.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog: a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42(Database issue):D1001-6.

Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. 2013;45:1238–1243.

Willrich MA, Murray DL, Snyder MR. Tumor necrosis factor inhibitors: clinical utility in autoimmune diseases. Transl Res. 2015;165(2):270-82.

Wise AL, Gyi L, Manolio TA. eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am J Hum Genet. 2013;92(5):643-7.

Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M. Dynamic trans-acting factor co-localization in human cells. Cell. 2013;155(3):713-24.

Xie X, Li F, Chen JW, Wang J. Risk of tuberculosis infection in anti-TNF-α biological therapy: from bench to bedside. J Microbiol Immunol Infect. 2014;47(4):268-74.

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43(6): 519–525.

Yin L, Chen X, Vicini P, Rup B, Hickling TP. Therapeutic outcomes, assessments, risk factors and mitigation efforts of immunogenicity of therapeutic protein products. Cell Immunol. 2015;295(2):118-126.

Zhang Z, Miteva MA, Wang L, Alexov E. Analyzing Effects of Naturally Occurring Missense Mutations. Comput Math Method in Medicine. 2012;2012:805827.

Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science.2008;322(5898):44.

Zhu LJ, Yang X, Yu XQ. Anti-TNF-α Therapies in Systemic Lupus Erythematosus. J Biomed Biotech. 2010;2010:465898. doi:10.1155/2010/465898.

Websites:

ACR aTNF data from ACR website:

http://www.rheumatology.org/I-Am-A/Patient-Caregiver/Treatments/Anti-TNF

ConsensusPathDB: http://consensuspathdb.org/

Cytoscape v3.1.1. : http://cytosca-pe.org/download.php

Drug Bank: http://www.drugbank.ca

ENCODE portal: http://encodeproject.org.

ENSEMBL portal: www.ensembl.org/Homo_sapiens/Info/Indexl

FDA aTNF data on FDA website:
http://www.fda.gov/drugs/drugsafety/postmarketdrugsafetyinformationforpatientsandproviders/ucm109340.htm

HAVANA project: http://vega.sanger.ac.uk/index.html

HLA Gene Family: http://ghr.nlm.nih.gov/geneFamily/hla

Human HLA Database: http://www.ebi.ac.uk/ipd/imgt/hla/stats.html

Human Proteome Map: http://www.humanproteomemap.org

miRBase: http://www.mirbase.org/

NCBI portal: http://www.ncbi.nlm.nih.gov/

NCI: http://pid.nci.nih.gov/

OMIM portal: http://www.ncbi.nlm.nih.gov/omim

PharmaGKB: https://www.pharmgkb.org/

PolyPhen-2: http://genetics.bwh.harvard.edu/pph2/

PubMed: http://www.ncbi.nlm.nih.gov/pubmed

Reactome: http://www.reactome.org/

STRING: http://string-db.org

TargetScan: http://www.targetscan.org/

7. TABLES AND FIGURES

Table 2. Missense GWAS AID SNPs characteristics: I part

(amino acid change caused by missSNPs; PolyPhen-2 evaluation of functional consequences; MAF)

No.	rs	Context	Gene name	Gene ID	Functional impact (PolyPhen-2)	1KG Project presence	MAF (minor allele count)	AA change	Disease	Comment
1	rs2476601	missense	PTPN22	26191	benign	1000 G	A=0.0274/137	V [Val] ⇒ A [Ala]	RA, CD (T1D)	same rs
2	rs4077515	missense	CARD9	64170	benign	1000 G	T=0.367/799	S [Ser] ⇒ N [Asn]	CD, UC, AS	same rs (has a near-gene SNP)
3	rs30187	missense	ERAP1	51752	damaging	1000 G	T=0.411/896	K [Lys] ⇒ R [Arg]	AS, Ps	same rs (has an intron rs)
4	rs11209026	missense	IL23R	149233	damaging	1000 G	A=0.033/72	R [Arg] ⇒ Q [Gln]	CD, AS, Ps, IBD, UC (asthma, T1D)	same rs
5	rs33980500	missense	TRAF3IP2	10758	damaging	1000 G	T=0.082/178	D [Asp] ⇒ N [Asn]	PsA,Ps, CD	same rs
6	rs2233434	missense	NFKBIE	4794	benign	1000 G	G=0.074/161	V [Val] ⇒ A [Ala]	RA	-
7	rs3197999	missense	MST1	4485	benign	1000 G	A=0.215/469	R [Arg] ⇒ C [Cys]	CD, UC	same rs
8	rs2241880	missense	ATG16L1	55054	benign	1000 G	G=0.389/847	T [Thr] ⇒ A [Ala]	CD, UC	same rs
9	rs3764147	missense	LACC1	144811	benign	1000 G	G=0.303/659	I [Ile] ⇒ V [Val]	CD (leprosy)	same rs
10	rs12720356	missense	TYK2	7297	damaging	1000 G	C=0.045/99	I [Ile] ⇒ S [Ser]	CD, Ps, PsA, UC, IBD, RA (T1D,T2D)	same rs
11	rs3184504	missense	SH2B3	10019	benign	1000 G	T=0.1474/738	W [Trp] ⇒ R [Arg]	RA (T1D, hypotyroidism, celiac disease)	same rs
12	rs2230926	missense	TNFAIP3	7128	benign	1000 G	G=0.124/270	F [Phe] ⇒ C [Cys]	RA (SLE, PS, UC, celiac d.), PsA	same rs (has intron rs)
13	rs2066847	frameshift	NOD2	64127	damaging	1000G	C=0.008/17	L [Leu] ⇒ P [Pro]	CD, IBD, Ps	same rs
14	rs20541	missense	IL13	3596	benign	1000G	A=0.268/584	Q [Gln] ⇒ R [Arg]	Ps (IgE)	same rs
15	rs2228145	missense	IL6R	3570	benign	1000G	C=0.319/695	D [Asp] ⇒ A [Ala]	RA (asthma, CRP, etc)	same rs
16	rs1063635	missense	MICA	100507436	no data	no	no data	R [Arg] ⇒ Q [Gln]	RA, Ps	only rs
17	rs3125734	missense	RTKN2	219790	benign	1000G	T=0.351/764	H [His] ⇒ R [Arg]	RA	only rs
18	rs2298428	missense	YDJC	150223	damaging	1000G	T=0.226/492	A [Ala] ⇒ T [Thr]	RA, CD (Celiac)	same rs
19	rs8192591	missense	NOTCH4	4855	damaging	1000G	T=0.0312/156	G [Gly] ⇒ S [Ser]	Ps (RA, scleroderma, SLE, T1D, T2D, MS, asthma, schizophrenia)	not same rs (3 miss and syn rss in high LD)
20	rs1801274	missense	FCGR2A	2212	benign	1000G	G=0.430/937	H [His] ⇒ R [Arg]	UC (mucokutaneous disease)	same rs
21	rs2305480	missense	GSDMB	55876	damaging	1000G	A=0.2867/1436	P [Pro] ⇒ S [Ser]	UC (asthma same; RA, T1D, CD different rs)	not same rs
22	rs5771069	missense	IL17REL	400935	benign	1000G	G=0.2208/1106	L [Leu] ⇒ P [Pro]	UC	-
23	rs3194051	missense	IL7R	3575	benign	1000G	G=0.2208/1106	I [Ile] ⇒ V [Val]	UC (T1D, MS)	different miss rs

Each AID missSNP is labeled with rs number, associated gene official symbol, NCBI gene identifier, presence in the 1000 Genome Project, minor allele frequency count (MAF), and the type of amino acid alteration caused by the missense mutation; AID where the missSNP is present; comment is given in regard to relation of missSNPs with other SNPs in the same gene if known.

Yellow field: the genes with damaging missSNP

Abbreviations : AS ankylosing spondylitis; CD Crohne disease; CRP C-reactive protein; IBD inflammatory bawl disease; MS multiple sclerosis; Ps psoriasis; PsA psoriatic arthritis; RA rheumatoid arthritis; SLE systemic lupus erythematosus; T1D type one diabetes;T2D type two diabetes; UC ulcerative colitis;

NFKBIE has an intronic 2b SNP rs3799963 in perfect LD with 2 missense SNPs rs2233434 and rs2233433

YDJC miss SNP rs 2298428 is in perfect LD with 3'UTR of UBE2L3 double hit rs7445 and rs7444, both in conserved regions; UBE2L3 could be also a functional gene

Table 3. Missense GWAS AID SNPs characteristics: II part

analysis of LD blocks by HaploReg v2 and RegulomeDB scoring

No	rs	Gene name	Gene ID	Variants in LD with the lead SNP	Genes in LD	Conserved region	Promoter histone marks	Enhanser histone marks	DNase	Proteins bound	eQTL tissues	Motifs changed	dbSNP functional annotation	Regulome DB score	Comment about lead vs functional SNP
1	rs2476601	PTPN22	26191	no other genes; lead SNP is functinal, well annotated and scored high	PTPN22	yes		HMEC	7 cell types	STAT3	13 eQTL results	Ets, NERF1a, PU.1	missense	2b	Lead and Functional SNP
2	rs4077515	CARD9	64170	two genes with two functional 1f SNPs in LD: CARD9 missSNP scored 1f, synSNP in LD; SNAPC4 has a cons missSNP scored 1f and several synSNPs, some conserv.; lead SNP is functinal but other functional SNPs might exist.	CARD9, SNAPC4	yes		HSMM, K562, NHLF			34 eQTL results		missense	1f	Lead and Functional SNP; other functional SNPs may exist in CARD9 and in SNAPC4
3	rs30187	ERAP1	51752	no other genes in LD; syn rs within the same gene.	ERAP1	yes			17 cell types		15 eQTL results	NF-kappaB, Zbtb3	missense	5	Lead and Functional SNP
4	rs11209026	IL23R	149233	no other genes in LD; intronic rss in LD less annoted; lead SNP is functional.	IL23R	yes					5 eQTL results	GR	missense	5	Lead and Functional SNP
5	rs33980500	TRAF3IP2	10758	no other genes in LD; lead SNP is functional.	TRAF3IP2, TRAF3IP2-AS1	yes		7 cell types	36 cell types		6 eQTL results	BCL,FXR, Maf, Pou1f1, STAT	missense	4	Lead and Functional SNP
6	rs2233434	NFKBIE	4794	2 miss SNPs within NFKBIE in perfect LD, both scored 4; several genes in LD: TCTE1, AARS2 with syn scored less than 5.	NFKBIE, TCTE1, AARS2	no	8 cell types	Huvec	6 cell types	CTCF, ZNF263	17 eQTL results		missense	4	Lead and Functional SNP; double hit
7	rs3197999	MST1	4485	in LD with APEH with syn cons SNP; BSN with miss and syn rss, scored less than 5; potentally lead SNP is not a functional SNP	MST1, APEH, BSN	yes					24 eQTL results	SMC3, TCF12, ZBRK1	missense	no data	Lead SNP most probably not Functional SNP
8	rs2241880	ATG16L1	55054	no other gene in LD; no other miss SNP; in perfect LD with cons intronic SNP of SCARNA5, with no scoring data.	ATG16L1	no					6 eQTL results	NRSF	missense	no data	Lead and Functional SNP
9	rs3764147	LACC1	144811	no other genes; high scoring 1f very likely to affect binding and expression.	LACC1	yes		HSMM	15 cell types		9 eQTL results	ATF3, GATA, Hand1, Mrg, Smad3	missense	1f	Lead and Functional SNP
10	rs12720356	TYK2	7297	no other genes; it is the only missSNP in TYK2 gene	TYK2	yes			8988T, Chorion		27 eQTL results		missense	5	Lead and Functional SNP
11	rs3184504	SH2B3	10019	ATXN2 gene in LD with several intronic SNPs scored low; lead SNP is a functional SNP.	SH2B3, ATXN2	yes		Huvec, GM12878			15 eQTL results	HES1, Mtf1	missense	3a	Lead and Functional SNP
12	rs2230926	TNFAIP3	7128	no other genes; lead SNP is a functional SNP	TNFAIP3	yes		M12878	NHEK		2 eQTL results	GR, Sin3Ak-20	missense	4	Lead and Functional SNP
13	rs2066847	NOD2	64127	no other genes in LD; lead SNP is a functional SNP	NOD2	no						BDP1, SIX5, SRF, p300	frameshift	5	Lead and Functional SNP

14	rs20541	IL13	3596	no other genes in LD; likely lead SNP is a functional SNP; no scoring data for other IL13 SNPs	IL13	no		M12878			8 eQTL results	RXRA	missense	3a	Lead and Functional SNP
15	rs2228145	IL6R	3570	no other genes in LD; two other intronic SNPs scored 2b likely affecting binding; lead SNP is not the only functional SNP.	IL6R	yes		HMEC, GM12878					missense	5	Lead SNP might not be the only Functional SNP
16	rs1063635	MICA	100507436	no other genes in LD; no other SNPs. Lead SNP scored 6; poor annotaion.	MICA	no					13 eQTL results	Sin3Ak-20	missense	6	Lead SNP might not be Functional SNP
17	rs3125734	RTKN2	219790	no other genes in LD; no other annotated SNPs.	RTKN2	yes					2 eQTL results	Maf	missense	5	Lead and Functional SNP
18	rs2298428	YDJC	150223	lead SNP scored 4; UBE2L3 in LD, has intronic SNPs, no miss SNPs; UBE2L3 3'-UTR SNPs scored 2b and 3a both cons; no QTL in LD.	YDJC, UBE2L3	yes	4 cell types		NB4, SK-N-SH RA	POL2, SIN3 AK20, CHD2, HEY1, SMC3	18 eQTL results	Irf, KAP1, Nanog, PRDM1, Pax-5, TATA	missense	4	Lead and but not Functional SNP; UBE2L3 gene is more important functionally with high scored SNPs.
19	rs8192591	NOTCH4	4855	lead Notch SNP scored 5; NOTCH4 has another miss and syn SNP in LD, both scored 1f, potentially both are functional, as the other NOTCH4 rss might be; EGFL8 in perfect LD and has a cons 3'-UTR SNP scored 2b; non-cons rs8192583 of GPSM3 scored 1f;	NOTCH4, EGFL8, GPSM3 (GPSM3 in lower LD than EGFL8)	no					15 eQTL results	BCL, NRSF	missense	5	Lead missSNP might not be the only functional SNP; more NOTCH4 functional SNPs may exist in LD; EGFL8 3'-UTR SNP might be functional.
20	rs1801274	FCGR2A	2212	no other genes in LD; several intronic, low scored.	FCGR2A	no					11 eQTL results		missense	no data	Lead and Functional SNP
21	rs2305480	GSDMB	55876	GSDMB has second missSNP ih perfect LD, both scored 1f; ZPBP2 in LD with one miss and several cons intronic SNPs (one is 1f).	GSDMB, ZPBP2	no			HepG2		23 eQTL results	BAF155, TAL1	missense	1f	Lead and functional SNP; two functional miss SNPs exist; double hit.
22	rs5771069	IL17REL	400935	PIM3 gene in LD with no miss or cons rs and all scored low; this lead SNP is scored 5 and might not be the only functional SNP.	IL17REL, PIM3	no		H1	LNCaP, GM 12892, Osteobl		17 eQTL results	BCL,Egr-1, Nanog, PRDM1, PU.1,STAT, TATA, YY1	missense	5	Lead and Functional SNP, potentally not the only one
23	rs3194051	IL17R	3575	no other missense SNP in LD, all SNPs scored low, 5 or less. CAPSL in LD (no function).	IL7R, CAPSL	no		GM12878, Huvec	Adult CD4, Th0, NHDF-Ad		9 eQTL results		missense	5	Lead and Functional SNP

Note: potentally damaging SNPs are shaded yellow

NFKBIE has intronic 2b SNP rs3799963 (not from any GWAS) in a perfect LD with 2 missense SNPs rs2233434 and rs2233433

Missense AID SNPs: each SNP is labeled with its rs identifier, and associated genes; whether the SNP is in high LD with other genes and SNPs, weather it is nested in conserved regions, or if it has any chromatin marks and functional annotations, and its RegulomeDB score. Also provided results of our evaluation based on all data if the missSNP is a functional in addition to being a lead GWAS SNP.

Table 4. Alternative coding genes at the missense AID GWAS SNP loci

Gene name	Gene ID	Function	Association with disease	KEGG pathway or disease		Comments	Role within Immune System	Selected
NOTCH4, EGFL8, GPSM3								
EGFL8	80864	unknown	Macular degeneration	none			unknown	
GPSM3	63940	unknown	Ps, SLE	none			unknown	
NOTCH4	4855	functions as a receptor for membrane bound ligands	RA, Ps, T1D, MS, macular degeneration, asthma, sceroderma systemic, etc.	hsa04320 hsa04330 hsa04919 hsa05206	Dorso-ventral axis formation Notch signaling pathway Thyroid hormone signaling pathway MicroRNAs in cancer	Notch family are Type 1 transmembrane proteins. The Notch signaling network is an evolutionarily conserved intercellular signaling pathway which regulates interactions between physically adjacent cells.	no	**x**
SH2B3, ATXN2								
SH2B3	10019	an intracellular adaptor protein involved in a range of signaling activities by growth factor and cytokine receptors; regulates B lymphopoiesis, megakaryopoiesis, and expansion of hematopoietic stem cells by constraining cytokine signals	T1D, RA, Coronary Artery Disease, Hypothyroidism, etc.	hsa04722 H00408	Neurotrophin signaling pathway T1D	Sh2b3 plays a regulatory role in the expansion of DCs and might influence inflammatory immune responses in peripheral lymphoid tissues. JAK2 is an established SH2B3 target	yes	**x**

ATXN2	6311	unknown, might participates in the regulation of RNA metabolism	Celiac disease, autoimmune diseases, etc.	H00063	Spinocerebellar ataxia (SCA)		no	
YDJC, UBE2L3								
YDJC	150223	unknown, but widely present	CD, Celiac disease, autoimmune diseases	none			no	
UBE2L3	7332	participant in ubiquitination, an important cellular mechanism for targeting abnormal or short-lived proteins for degradation; widely present.	CD, SLE	hsa04120	Ubiquitin mediated proteolysis	This gene encodes a member of the E2 ubiquitin-conjugating enzyme family; demonstrated to participate in the ubiquitination of p53, c-Fos, and the NF-kB precursor p105 in vitro.	no	**x**
GSDMB, ZPBP2								
GSDMB	55876	implicated in the regulation of apoptosis in epithelial cells, and is linked to cancer	UC, T1D, RA, CD, asthma,	none		a member of the gasdermin-domain containing protein family	no	**x**
ZPBP2	124626	zona pellucida binding protein 2	UC, CD, RA, T1D	none		no annotaion	no	
IL17REL, PIM3								
IL17REL	132014	member of IL-17 receptors; signal transduction pathway mediated by IL-17 is still poorly defined; IL-17 receptors activate noncanonical signal transduction pathways	UC, also Ps, RA, uveitis, etc.	ko04060	Cytokine-cytokine receptor interaction	Activity of this protein is important in the immune response to bacterial and fungal pathogens. Encoded protein signals to downstream components of the MAPK pathway, signals through Act1; activates the classical NF-κB pathway. There are at least 5 IL-17R proteins	yes	**x**

PIM3	415116	Pim-3 proto-oncogene, serine/threonine kinase	none	none		no annotaion	unknown	
IL7R, CAPSL								
CAPSL	133690	calcyphosine-like	T1D, Liver Cirrhosis	none		unknown	no	
IL7R	3575	protein encoded by this gene is a receptor for interleukine 7; shown to play a critical role in the V(D)J recombination during lymphocyte development	UC, T1D, MS, neoplasms	hsa04060 hsa04068 hsa04151 hsa04630 hsa04640 hsa05340	Cytokine-cytokineR interaction FoxO signaling pathway PI3K-Akt signaling pathway Jak-STAT signaling pathway Hematopoietic cell lineage Primary immunodeficiency	control the accessibility of the TCR gamma locus by STAT5 and histone acetylation; also blocking apoptosis is an essential function of this protein during differentiation and activation of T lymphocytes; may be associated with the pathogenesis of the severe combined immunodeficiency (SCID).	yes	**x**
MST1, APEH, BSN								
MST1	4485	encoded protein contains four kringle domains and a serine protease domain, similar to that found in hepatic growth factor, but no proteolytic activity	UC, CD	none		similar to macrophage-stimulating 1 receptor	no	**x**
APEH	327	encodes the enzyme acylpeptide hydrolase, which catalyzes the hydrolysis of the terminal acetylated amino acid preferentially from small acetylated peptides.	UC	none		acylaminoacyl-peptide hydrolase	no	

BSN	8927	gene is expressed primarily in neurons in the brain; encodes a scaffolding protein involved in organizing the presynaptic cytoskeleton	UC, CD	none		bassoon presynaptic cytomatrix protein	no	
NFKBIE, TCTE1, AARS2								
NFKBIE	4794	The protein encoded by this gene binds to components of NF-kappa-B, trapping the complex in the cytoplasm and preventing it from activating genes in the nucleus.	RA	yes	many KEGG pathways	Phosphorylation of the encoded protein targets it for destruction by the ubiquitin pathway, which activates NF-kappa-B by making it available to translocate to the nucleus		**x**
TCTE1	202500	t-complex-associated-testis-expr	none	none		no annotaion		
AARS2	57505	the encoded protein is a mitochondrial enzyme that specifically aminoacylates alanyl-tRNA	MI	none		Mutations in this gene are a cause of combined oxidative phosphorylation deficiency 8.		
CARD9, SNAPC4								
CARD9	64170	this gene is a member of the CARD protein family, which is defined by the presence of a characteristic caspase-associated recruitment domain (CARD)	UC, CD, AS	hsa04621 hsa05152	NOD-like receptor signaling pathway Tuberculosis	This protein was identified by its selective association with the CARD domain of BCL10, a postive regulator of apoptosis and NF-kappaB activation,		**x**
SNAPC4	6621	snRNA-activating protein complex subunit 4	none	none		Transcription factors: potental role in relation to GWAS SNP genes does not exist		

Coding genes associated with missense AID SNPs: name and ID of coding genes, function if known, association with diseases, participation in canonical KEGG pathways and role in the immune system.

Table 5. Characteristics of highest Regulome DB scored GWAS AID ncSNPs

No	AID	db SNP ID	Score	Context	Gene name	GWAS p value	Source	PubMed	Comments	Genes in high LD (> 0.8) including ncRNAs
1	**IBD**	rs8049439	**1b**	intron	ATXN2L	2.000 x 10-9	NHGRI	19915574	not cons, in high LD (0.8) with another rs cons ATXN2L intronic; ataxin 2-like; in LD with TUFM intronic rs	TUFM in high LD (translation elongation factor) ATXN2L microRNA 4721 in high LD
2	**UC**	rs798502	**1b**	intron	GNA12	3.000 x 10-15	NHGRI	21297633	intronic non cons SNP is in LD (0.8) with cons rss in GNA12	GNA12
3		rs11676348	**1b**	intergenic	CXCR2, CXCR1	1.000 x 10-10	NHGRI	21297633	in LD with CXCR2 gene; in high LD with CXCR2 syn cons SNP; in LD with intragenic cons CXCR1 rs that is site for STAT3 binding	CXCR1, CXCR2 (chemokine (C-X-C motif) receptor 1 and 2)
4		rs9263739	**1f**	intron	CCHCR1	4.000 x 10-67	NHGRI	19915573	not cons; in total LD (1) with 2 missense SNPs of PSORS1C1 gene; in high LD for PSORC1C2 and C3 (psoriasis susceptibility 1 candidate 2 and 3)	PSORS1C1 (Gene ID: 170679), 2 (Gene ID: 170680) and 3 (PSORS1C3 GeneID: 100130889 is ncRNA gene)
5		rs4728142	**1f**	intergenic	KCP, IRF5	2.000 x 10-8	NHGRI	21297633	in LD with IRF5 intronicSNPs	IRF5
6		rs9858542	**1f**	intron	BSN	7.000 x 10-9	NHGRI	19915572	in perfect LD with miss, syn, and 3'UTR BSN and miss MST1	MST1
7		rs10781499	**1f**	intron	CARD9	3.000 x 10-19	NHGRI	21297633	in high LD (0.98) with missense SNP in CARD9 already accounted for; in high LD with cons miss SNP within SNAPC4	CARD9
8		rs734999	**1f**	intergenic	TNFRSF14, C1orf93	3.000 x 10-9	NHGRI	21297633	in LD with miss SNP in TNFRSF14 a member of the TNFreceptor superfamily	TNFRSF14
9		rs8067378	**1f**	intergenic	ZPBP2, GSDMB	1.000 x 10-7	NHGRI	20228799	in high or perfect LD with a cons miss SNPs of ZPBP2 and 2 miss SNPs of GSDMB; both genes have no known function	ZPBP2 and GSDMB
10		rs907611	**2a**	nearGene-5	LSP1	1.000 x 10-10	NHGRI	21297633	in perfect LD with 5'-UTR of LSP1	LPS1
11		rs1297265	**2a**	intergenic	NRIP1, CYCSP42	7.000 x 10-13	NHGRI	21297633	not cons, no genes in LD; only several cons regions in LD.	none

12		rs3024505	**2b**	intergenic	RPS14P1, IL10	6.000 x 10-17	NHGRI	21297633	3' of IL10; this rs also linked with T1D, IBD, CD and UC by PhenGenI	IL10
13		rs3806308	**2b**	near gene-5	RNF186	7.000 x 10-9	NHGRI	19122664	not cons, one of several UC linked rs totally intergenic, no genes in LD	none
14		rs3024493	**2b**	intron	IL10	1.000 x 10-12	NHGRI	20228798	not cons; on other genes in LD	IL10
15		rs6451493	**3a**	intergenic	DAB2, PTGER4	3.000 x 10-9	NHGRI	21297633	no coding genes in LD; a novel miRNA in high LD	none
16	**RA**	rs660895	**1f**	intergenic	HLA-DRB1, HLA-DQA1	1.000 x 10-108	NHGRI	17804836	no cons (HLA complex); no other genes; no ncRNA	HLA
17		rs3093023	**1f**	intron	CCR6	2.000 x 10-11	NHGRI	20453842	not cons; in high LD with many SNPs intronic to chemokine (C-C motif) receptor 6 with OMIM for RA; participate in C-C pathways; no ncRNa	CCR6
18		rs805297	**1f**	nearGene-5	APOM	3.000 x 10-10	NHGRI	21844665	no cons, LD high with 3 SNPs in cons regions of 3 genes BAG6 intronic rs, PRRC2A syn rs, CSNK2B 3' UTR cons rs; and two intronic rs in two genes LY6G5B, LY6G6F. Note: BAG4 or SODD athanogene is tumor necrosis factor receptor superfamily member 1A (TNFRSF1A ID 7132) inhibitor/mpdulator)	BAG6 is BCL2-associated athanogene PRRC2A LY6 cluster APOM CSNK2B
19		rs3781913	**1f**	intron	PDE2A	6.000 x 10-10	NHGRI	22446963	not cons, intronic in LD with cons SNP in the same gene (phosphodiesterase 2A, cGMP-stimulated) and two nc RNA	microRNA 139; Gene ID: 406931 PDE2A ncRNA uncharacterized LOC102724448
20		rs4810485	**1f**	intron	CD40	3.000 x 10-9	NHGRI	20453842	not cons; in perfect LD with cons 5'-UTR rs and a cons near Gene rs in CD40 (TNF receptor superfamily member 5) and intronic rs of NCOA5, nuclear receptor coactivator 5	NCOA5 CD40
21		rs881375	**2b**	intergenic	PHF19, TRAF1	4.000 x 10-8	NHGRI	19503088	highly cons, 3' nearGene of TRAF1; in perfect LD with many intronic, syn and 3'-UTR of TRAF1 (some are cons)	TRAF1
22		rs874040	**2b**	intergenic	C4orf52, RBPJ	1.000 x 10-16	NHGRI	20453842	recombination signal binding protein for immunoglobulin kappa J region; Gene ID: 3516. The protein encoded by this gene is a transcriptional regulator important in the Notch signaling pathway.	RBPJ

23		rs6859219	**2b**	intron	ANKRD55	1.000 x 10-11	NHGRI	20453842	non cons intronic; this same rs6859219 is also multiple sclerosis risk SNP	ANKRD55
24		rs26232	**2b**	intron	C5orf30	4.000 x 10-8	NHGRI	20453842	coding region, no genes	none
25		rs12831974	**2b**	intron	THRADE	6.000 x 10-7	NHGRI	21452313	thyrotropin-releasing hormone degrading enzyme; an extracellular peptidase; no other genes or SNPs in LD	THRADE
26		rs10488631	**3a**	nearGene-3	IRF5, TNPO3	4.000 x 10-11	NHGRI	20453842	not cons; 3' of TNOP3 (transportin 3, tRNA import factor is a nuclear import receptor for serine/arginine-rich (SR) proteins); in perfect LD with 3' UTR of IRF5 (encodes a member of the interferon regulatory transcription factor (IRF) family).	IRF5, TNPO3
27		rs3783637	**2b**	intron	GCH1	2.000 x 10-7	NHGRI	20453842	GTP cyclohydrolase I; noother SNPs in LD	GCH1
28		rs3087243	**3a**	nearGene-3	CTLA4	1.000 x 10-8	NHGRI	20453842	non cons, in LD with many rs in 3' nearGene of CTLA4; no other genes in LD	CTLA4
29		rs26232	**2b**	intron	C5orf30	4.000 x 10-8	NHGRI	20453842	coding region, no genes	none
30		rs12831974	**2b**	intron	THRADE	6.000 x 10-7	NHGRI	21452313	thyrotropin-releasing hormone degrading enzyme; an extracellular peptidase. no other genes or SNPs in LD	THRADE
31		rs16906916	**3a**	intergenic	SV2B, TRNAY16P	5.000 x 10-7	NHGRI	22491018	Novel lincRNA; SLCO3A1 in very low LD	none
32	**CD**	rs181359	**1f**	intron	UBE2L3	5.000 x 10-16	NHGRI	21102463	not cons; in perfect LD with intronic SNPs of UBE2L3 only	UBE2L3
33		rs2549794	**1f**	intron	ERAP2	1.000 x 10-10	NHGRI	21102463	not cons; in perfect LD with intronic ERAP2	ERAP2
34		rs6596075	**1f**	intergenic	SLC22A5, C5orf56	3.000 x 10-7	NHGRI	17554300	not cons; in LD with cons 3'-UTR of SLC22A5 and few intronic SNP of the same gene	SLC22A5
35		rs694739	**1f**	intergenic	PRDX5, CCDC88B	6.000 x 10-10	NHGRI	21102463	not cons; in strong LD with miss, 5'-UTR and syn of gene CCDC88B;	CCDC88B
36		rs713875	**1f**	intergenic	RPS3AP51, LIF 22	7.000 x 10-12	NHGRI	21102463	not cons; in LD with HORMAD2 intergenic SNPs; novel antisense RP3-438O4.4 ; not in LD with ribosomal protein S3a pseudogene 51	HORMAD2
37		rs9258260	**1f**	intergenic	IFITM4P, 3.8-1.5	2.000 x 10-10	NHGRI	22412388	not cons;intergenic, no functional genes in LD except interferon induced transmembrane protein 4 pseudogene; non-coding RNAs (ncRNAs); HLA-F antisense RNA 1	none
38		rs9858542	**1f**	intron	BSN	7.000 x 10-9	NHGRI	19915572	in perfect LD with miss, syn, and 3'UTR BSN and miss MST1	MST1

39		rs17293632	**2a**	intron	SMAD3	3.000 x 10-19	NHGRI	21102463	not cons, only SNAD3 intronic SNPs in LD	SMAD3
40		rs2188962	**2a**	intron	C5orf56	2.000 x 10-18	NHGRI	18587394	not conservative; in high LD with miss SNP in SLC22A4 (6583); in LD with cons SNP in 5'UTR of IRF1 gene.	SLC22A4: solute carrier family 22 organic cation transporter IRF1: interferon regulatory factor 1, a transcription factor two ncRNAs: uncharacterized LOC102723741; uncharacterized LOC101927732
41		rs4263839	**2b**	intron	TNFSF15	3.000 x 10-10	NHGRI	18587394	not cons; only TNFSF15 intronic SNP in LD	TNFSF15
42		rs3024505	**2b**	intergenic	RPS14P1, IL10	2.000 x 10-14	NHGRI	21102463	3' -UTR and intronic rs in IL10 strong LD (>0.8); not cons, no LD for cons SNPs	IL10
43		rs4409764	**3a**	intergenic	GOT1, NKX2-3	2.000 x 10-20	NHGRI	21102463	highly conservative rs in the region of 3' ncRNA; in perfect LD with syn SNP of NKX2-3, NK2 homeobox 3 transcription factor and a few intronic SNP of the same gene	NKX2-3 long intergenic non-protein coding RNA 1475
	Ps & PsA									
44		rs9267673	**1f**	intron	ZBTB12, C2	2.300 x 10-33	dbGaP	phs000019	not cons, intronic of C2 in HaploRegv2; in LD with intronic SNP of SKIV2L	C2 (complement component, an serum glycoprotein of the classical pathway). SKIV2L: superkiller viralicidic activity 2-like
45		rs3130955	**1f**	intergenic	HCG22, C6orf15	1.312 x 10-20	dbGaP	phs000019	RNA, U6 small nuclear 1133, pseudogene	PSORS1C1 psoriasis susceptibility 1 candidate 1 is the closeset functional gene but in a very weak LD (r2 < 0.2) RNA, U6 small nuclear 1133, pseudogene HCG22 HLA complex group 22 (ID: 285834)
46		rs8192583	**1f**	UTR-5	GPSM3	2.250 x 10-20	dbGaP	phs000019	syn non cons SNP of G-protein signaling modulator 3; in high LD with syn cons SNP of NOTCH4	NOTCH4
47		rs8192583	**1f**	intronic	NOTCH4	2.250 x 10-20	dbGaP	phs000019	NOTCH4 (acually both is true); inperfect LD with each other and gene PPT2, an thioesterase	GPSM3

										HLA-C (Gene ID: 3107)	
48		rs2243868	**1f**	intergenic	WASF5P, HLA-B (3 rs on chr.6 all scored 1f)	2.798 x 10-19	dbGaP	phs000019	not cons; WASF5P is pseudogene (Chr. 6) of WASF3 (Chr.13); belongs to the family of genes encoding Wiskott-Aldrich syndrome (WAS) proteins. Wiskott-Aldrich syndrome is a disease of the immune system (thrombocytopenia); 3' of lincRNA gene XXbac-BPG248L24.13; no coding genes in LD;HLA-C rs in 5' UTR in perfect LD and with several missSNPs in very low LD (0.5)	WASF5P	HLA-B
49		rs2395471	**1f**	nearGene-5	HLA-C	8.836 x 10-18	dbGaP	phs000019	rs not cons in 5' HLA-C region; no other structures in LD up to 0.2	HLA-C	
50		rs3094187	**1f**	UTR-5	TCF19	3.725 x 10-13	dbGaP	phs000019	TCF19, transcription factor 19	TCF19	
51		rs8365	**1f**	UTR-3	RNF5	1.094 x 10-17	dbGaP	phs000019	ring finger protein 5, E3 ubiquitin protein ligase	RNF5	
52		rs176095	**1f**	nearGene-5	PBX2	2.183 x 10-12	dbGaP	phs000019	G-protein signaling modulator 3; high LD with pre-B-cell leukemia homeobox 2, an TF	PBX2	
53		rs1265086	**1f**	nearGene-3	CCHCR1	2.689 x 10-11	dbGaP	phs000019	not cons; in total LD (1) with 2 missense SNPs of PSORS1C1 gene; in high LD for PSORC1C2 and C3 (psoriasis susceptibility 1 candidate 2 and 3)	PSORS1C1 (Gene ID: 170679), 2 (Gene ID: 170680) and 3 (PSORS1C3 GeneID: 100130889 is ncRNA gene)	
54		rs2239518	**1f**	intron	DDR1	3.454 x 10-12	dbGaP	phs000019	discoidin domain receptor tyrosine kinase 1; involved in the regulation of cell growth, differentiation and metabolism. in high LD with miss SNP of VARS2, valyl-tRNA synthetase 2, mitochondrial	DDR1, VARS2	
55		rs3131043	**1f**	intergenic	IER3, DDR1	1.814 x 10-11	dbGaP	phs000019	HLA complex group 20 (non-protein coding); has two SNPs scored 2b within gene LINC00243, long intergenic non-protein coding RNA 243; same location has several SNPs all conecte with HCG20	HCG20 and LINC00243	
56		rs3130573	**1f**	intron	PSORS1C1	6.212 x 10-15	dbGaP	phs000019	no other SNPs in LD	PSORS1C1	
57		rs9262492	**2b**	intergenic	MUC21, HCG22	2.993 x 10-11	dbGaP	phs000019	in LD with MUC22, no SNPs, no function; HLA complex close but not in perfect LD	MUC22	HCG22, HLA complex
58		rs12580100	**3a**	intergenic	RPS26, ERBB3	1.000 x 10-7	NHGRI	20953189	no genes in LD	none	
59	**AS**	rs13210693	**2a**	intergenic	FLJ37396, CCDC162	9.000 x 10-7	NHGRI	22138694	no genes in LD; several cons missSNPs in C6orf183, known polymorphic pseudogene	none	

Characteristics of the RegulomeDB highest scored ncSNPs: AID association, rs identifier, context, provisionally assigned gene with official gene symbols, GWAS p value, info source and PubMed publication identifier; weather it is nested in conserved regions, and whether the SNP is in LD with other rs or genes labeled with NCBI gene identifiers.

Table 6. Non-coding RNA genes in LD with top intergenic GWAS AID SNPs

No	rs #	Context	Gene closeset to SNP location	Cons	Genes in LD (by HaploReg v2)	No of ncRNA genes	ncRNA Gene ID	Name of ncRNA genes	miRNA gene present in a SNP region	any other type of RNA gene is present	no ncRNA gene present	Regulome DB score for the rs *
1	rs6457620	intergenic	HLA-DQB1, HLA-DQA2	no	23kb-25kb 5' of HLA-DQB1; 21kb 5' of XXbac-BPG254F23.7	1	100616218	microRNA 3135b	yes			**1f**
2	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	no	16kb-28kb 3' of HLA-DRA	0					none	
3	rs660895	intergenic	HLA-DRB1, HLA-DQA1	no	19kb-14kb 5' of HLA-DQA1	1	100616218	microRNA 3135b	yes			**1f**
4	rs6457617	intergenic	HLA-DQB1, HLA-DQA2	no	23kb-25kb 5' of HLA-DQB1; 25kb-21kb 5' of XXbac-BPG254F23.7	1	100616218	microRNA 3135b	yes			
5	rs13192471	intergenic	HLA-DQB1, HLA-DQA2	no	21kb-12kb 5' of XXbac-BPG254F23.7	1	100616218	microRNA 3135b	yes			
6	rs9272219	intergenic	HLA-DRB1, HLA-DQA1	no	HLA-DQA1	1	100616218	microRNA 3135b	yes			**1f**
7	rs6679677	intergenic	RSBN1, PHTF1	no	PTPN22 in high LD with miss SNP, 3' of RP11-129J12.2	1	101927324	LINC01475, long intergenic non-protein coding RNA 1475		yes		
8	rs9296015	intergenic	NOTCH4, C6orf10	no	2.1kb-4.8kb 3' of XXbac-BPG154L12.4; 9.1kb-1.9kb 5' of XXbac-BPG154L12.4;	1	102466190	microRNA 6721	yes			
9	rs615672	intergenic	HLA-DRB1, HLA-DQA1	no	16kb-19kb 5' of HLA-DRB1	0					none	
10	rs11742570	intergenic	DAB2, PTGER4	medium	2.8kb-120kb 3' of AC108105.1	2	619564	SNORD72 small nucleolar RNA, C/D box72		yes		
							102467077	LINC00603 long intergenic non-protein coding RNA 603		yes		

11	rs4613763	intergenic	DAB2, PTGER4	none	72kb 3' of AC108105.1	2	619564	SNORD72 small nucleolar RNA, C/D box 72		yes		
							102467077	LINC00603 long intergenic non-protein coding RNA 603		yes		
12	rs2413583	intergenic	PDGFB, RPL3	in strong LD with conservative SNPs	2.7kb 5' of AL031590.1	4	116936	RNU86 RNA, U86 small nucleolar		yes		
							26807	SNORD43 small nucleolar RNA, C/D box 43		yes		
							116937	SNORD83A small nucleolar RNA, C/D box 83A		yes		
							116938	SNORD83B small nucleolar RNA, C/D box 83B		yes		
13	rs10761659	intergenic	ZNF365	none	14kb 3' of ZNF365	1	uncharacterized LOC283045	ncRNA		yes		
14	rs4409764	intergenic	GOT1, NKX2-3	high	NKX2-3; LINC01475	1	101927324	LINC01475 long intergenic non-protein coding RNA 1475		yes		**3a**
15	rs10995271	intergenic	ZNF365	none	ZNF365	0	NA	none			none	
16	rs7714584	intergenic	IRGM, ZNF300	none	IRGM, a member of the p47 immunity-related GTPase family (Gene ID 345611)	0	NA	none			none	
17	rs9292777	intergenic	DAB2, PTGER4	high	117kb 3' of AC108105.1 a novel miRNA and PTGER4	1	102467077		yes			
18	rs6651252	intergenic	LINC00824	high cons	LINC00824	7	5820	PVT1 Pvt1 oncogene (non-protein coding)		yes		
							100302281	microRNA 1208	yes			
							100302175	microRNA 1207	yes			
							100302170	microRNA 1206	yes			
							100302161	microRNA 1205	yes			
							102723587	uncharacterized LOC102723587		yes		
							101927774	LINCRNA824 long intergenic non-protein coding RNA 824		yes		
19	rs2542151	intergenic	PSMG2, PTPN2	none	4kb 5' of RP11-973H7.1	0		none			none	
20	rs7743761	intergenic	DHFRP2, HLA-S	none	1.8kb 5' of U6	1	102465537	microRNA 6891	yes			

21	rs4349859	intergenic	HLA-S, MICA	none	1.8kb 5' of MICA	1	101929111	long intergenic non-protein coding RNA 1149		yes		
22	rs2242944	intergenic	FLJ45139, RPL2 3AP12	none	64kb 5' of AF064858.10	1	101928435	LOC101928435, ncRNA uncharactarized		yes		**2b**
23	rs10865331	intergenic	B3GNT2, TMEM 17	none	59kb 3' of snoU13 and B3GNT2	1	100847087	microRNA 5192	yes	yes		
24	rs11616188	intergenic	LTBR, RPL31P10	none	LTBR and 328bp 5' of RP1-102E24.8 a novel linc RNA	1	RP1-102E24.8 (Vega gene)	linc RNA		yes		
25	rs378108	intergenic	FLJ45139, RPL2 3AP12	medium	68kb 5' of AF064858.10, a novel linc RNA	1	101928435	LOC101928435, ncRNA uncharactarized		yes		
26	rs11249215	intergenic	RUNX3, SYF2	none	within RP11-84D1.2 (Vega gene), a novel linc RNA	1	100616365	microRNA 4425	yes			
27	rs4552569	intergenic	RPL13AP14, EDIL3	none	63kb 3' of EDIL3	0					none	
28	rs6556416	intergenic	IL12B, ADRA1B	none	29kb 3' of AC008697.1, a novel lincRNA	1	100873889	RNA, U4atac small nuclear 2, pseudogene		yes		
29	rs9267673	intergenic	ZBTB12, C2	no	intronic SNP of C2	0					no data	**1f**
30	rs10484554	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
31	rs9468933	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
32	rs2894207	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
33	rs9380237	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
34	rs2524163	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
35	rs2243868	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		**1f**
36	rs2853923	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
37	rs9380240	intergenic	WASF5P, HLA-B	no	HLA-C, a novel lincRNA, WASF5P, HLA-B and USP8P1	1	10246553	microRNA 6891	yes	yes		
38	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	no	17kb 3' of HLA-DRA	0					none	
39	rs6426833	intergenic	RNF186, OTUD3	no	RNF186 and 26kb 3' of RP11-91K11.2 Novel antisense; TMCO4	0					none	**2c**
40	rs10758669	intergenic	RCL1, JAK2	no	3.4kb 5' of JAK2	1	406894	microRNA 101-2	yes			

41	rs9268877	intergenic	HLA-DRB9, HLA-DRB5	no	18kb 3' of HLA-DRA	0					none	
42	rs2836878	intergenic	FLJ45139, RPL23AP12	high	PSMG1 and 64kb 5' of AF064858.10 a novel linc RNA	0	AF064858.10 (Vega gene)	a novel linc RNA		yes		
43	rs2395185	intergenic	HLA-DRB9, HLA-DRB5	no	20kb 3' of HLA-DRA	0					none	
44	rs6871626	intergenic	IL12B, ADRA1B	no	37kb 3' of AC008697.1	1	285627	uncharacterized LOC285627		yes		
45	rs6584283	intergenic	GOT1, NKX2-3	medium	RP11-129J12.2, a novel linc RNA	1	101927324	LINC1475, long intergenic non-protein coding RNA 1475		yes		
46	rs6017342	intergenic	HNF4A, RPL37AP1	no	5kb 3' of HNF4A	3	100500813	microRNA 3646	yes			
							101927219	HNF4A-AS1, HNF4A antisense RNA 1		yes		
							101927242	LINC01430, long intergenic non-protein coding RNA 1430		yes		
47	rs2006996	intergenic	TNFSF15, TNFSF8	no	24kb 5' of TNFSF15	0	no LD with any other gene				none	
48	rs2836878	intergenic	FLJ45139, RPL23AP12	no	PSMG1 and 64kb 5' of AF064858.10	1	102724740	LOC102724740, ncRNA uncharacterized LOC102724740		yes		
49	rs9271366	intergenic	HLA-DRB1, HLA-DQA1	no	9.1kb 5' of HLA-DQA1	1	100616218	microRNA 3135b	yes			
50	rs10500264	intergenic	SLC7A10, CEBPA	no	34kb 5' of SLC7A10	1	80054	CEBPA-AS1, CEBPA antisense RNA 1		yes		
51	rs11209032	intergenic	IL23R, IL12RB2	no	IL23R, and 6.7kb 5' of U4atac	1	102724481	RNA, U4atac small nuclear 4, pseudogene		yes		
52	rs477515	intergenic	HLA-DRB1, HLA-DQA1	no	12kb 5' of HLA-DRB1	1	100616218	microRNA 3135b	yes			
53	rs8054797	intergenic	BRD7, NKD1	no	BRD7 and 5.6kb 3' of RP11-21B23.2	1	102465462	microRNA 6771	yes	yes		
Total nc RNA genes in LD:									**27/53**	**36/53**	**14/53**	**8 of 53 rs**

* If no score is given, the score was more than 3: 4,5,6 or no data existed at the time of data collection

ncSNP and ncRNAs: official rs identifier for each ncSNP, genomic context, originally assigned gene with official gene symbols, weather it is nested in conserved regions, name of genes found in LD, NCBI gene identifier and number of RNA genes, whether miRNA or other type of RNA genes are present or not, and RegulomeDB score for each SNP

Table 7a. miRNA genes linked with AID GWAS ncSNPs found in TargetScanHumanDB

No	miRNA in high LD with GWAS AID SNPs	Found among TargetScan Human DB	Number of targeted conseved sites	Targeting AID SNP genes
1	MIR101-2	yes	803 transcripts with conserved sites, with a total of 915 conserved sites and 337 poorly conserved sites.	no
2	MIR1205	yes	426 transcripts with conserved sites, with a total of 448 conserved sites and 329 poorly conserved sites.	no
3	MIR1206	yes	215 transcripts with conserved sites, with a total of 225 conserved sites and 93 poorly conserved sites.	no
4	MIR1207	no		no
5	MIR1208	yes	322 transcripts with conserved sites, with a total of 337 conserved sites and 146 poorly conserved sites.	no
6	MIR3135B	yes	63 transcripts with conserved sites, with a total of 63 conserved sites and 10 poorly conserved sites.	no
7	MIR3646	yes		no
8	MIR4425	yes	1141 transcripts with conserved sites, with a total of 1351 conserved sites and 1387 poorly conserved sites.	no
9	MIR5192	no		no
10	MIR6721	no		no
11	MIR6771	no		no
12	MIR6891	no		no

Official name of miRNA genes, whether it was found in the miRNA database, number of miRNA targets and whether it targets any AID SNP harboring genes

Table 7b. Micro RNAs targeting 3'-UTR of the AID GWAS SNP genes

No	Gene name	Name	Conservation level of miRNA family
1	PTPN22	hsa-miR-3646	conserved
2	CARD9	none	
3	ERAP1	has-miR-143	conserved
		hsa-miR-4770	conserved
4	IL23R	hsa-miR-4729	conserved
5	TRAF3IP2	hsa-miR-4795	conserved
		hsa-miR-548n	conserved
6	NFKBIE	hsa-miR-2115	conserved
		hsa-miR-1184	conserved
		hsa-miR-4667-3p	conserved
		hsa-miR-3180-5p	conserved
		hsa-miR-1264	conserved
		hsa-miR-3611	conserved
		hsa-miR-2355-5p	conserved
		hsa-miR-3591-5p	conserved
7	MST1	hsa-miR-654-3p	conserved
		hsa-miR-3916	conserved
		hsa-miR-3125	conserved
		hsa-miR-583	conserved
8	ATG16L1	hsa-miR-142-3p	conserved
9	LACC1	no data	
10	TYK2	miR-124ab	broadly conserved among vertebrates
		hsa-miR-506	conserved
11	SH2B3	hsa-miR-181abcd	conserved
		hsa-miR-4262	conserved

12	TNFAIP3	miR-23	broadly conserved among vertebrates
		miR-125	broadly conserved among vertebrates
		miR-29	broadly conserved among vertebrates
13	NOD2	miR-122	broadly conserved among vertebrates
		miR-30	broadly conserved among vertebrates
14	IL13	has-miR-155	conserved
15	IL6R	has-miR-9	conserved
		has-miR-124	conserved
16	MICA	none	
17	RTKN2	hsa-miR-204	conserved
		hsa-miR-211	conserved
18	YDJC	hsa-miR-1184	conserved
		hsa-miR-4477a	conserved
19	NOTCH4	hsa-miR-3945	conserved
		hsa-miR-4265	conserved
		hsa-miR-4322	conserved
		hsa-miR-4296	conserved
		hsa-miR-607	conserved
20	FCGR2A	hsa-miR-4490	conserved
		hsa-miR-3691-5p	conserved
		hsa-miR-4752	conserved
		hsa-miR-3911	conserved
21	TNF	miR-19	broadly conserved among vertebrates

Name of the missSNP harboring gene which 3'-UTR was inspected for miRNA, official name of miRNA genes in their 3'-UTR regions, and miRNAs level of conservation

Table 7c. GWAS AID SNPs found in the UTR-3' regions of targeted genes

(found by PhenGenI, analized by HaploReg and RegulomeDB)

No	AID	rs	Context	Gene	Location		P-value	Source	ncRNA in LD	Regulome DB score
1	Ps	rs2395029	UTR-3	HCP5	6	: 31,431,780	2.000 x 10-26	NHGRI	lincRNA 1149	2b
2	Ps	rs8365	UTR-3	RNF5	6	: 32,148,403	1.094 x 10-17	dbGaP	none	1f
3	Ps	rs2240803	UTR-3	DPCR1	6	: 30,920,957	8.193 x 10-13	dbGaP	novel misc RNA, Y RNA, ncRNA	5
4	IBD	rs10889677	UTR-3	IL23R	1	: 67,725,120	9.037 x 10-11	dbGaP	none	no data
5	UC	rs2297441	UTR-3	RTEL1	20	: 62,327,582	2.000 x 10-10	NHGRI	none	4
6	UC	rs10889677	UTR-3	IL23R	1	: 67,725,120	1.000 x 10-8	NHGRI	none	no data
7	CD	rs504963	UTR-3	FUT2	19	: 49,208,865	2.000 x 10-8	NHGRI	none, in LD with nonsence SNP	5
8	RA	rs1329568	UTR-3	LOC100130458	9	: 37,037,976	8.000 x 10-7	NHGRI	none	4

Name of the missSNP harboring gene which 3'-UTR was inspected for miRNA, official name of miRNA genes in their 3'-UTR regions, and miRNAs level of conservation

Table 8. Genes in missSNP gene set and ncSNP gene set

missSNP set	ncSNP set
PTPN22	ANKRD55
CARD9	APOM
ERAP1	ATXN2L
IL23R	BAG6
TRAF3IP2	C2
NFKBIE	CCR6
MST1	CD40
ATG16L1	CTLA4
LACC1	CXCR1 & 2
TYK2	CCDC88B
SH2B3	DDR1
TNFAIP3	ERAP2
NOD2	GNA12
IL13	GPSM3
IL6R	HORMAD2
MICA	IL10
RTKN2	IRF1
YDJC	IRF5
NOTCH4	LPS1
FCGR2A	NKX2-3
GSDMB	PBX2
IL17REL	PDE2A
IL7R	PSORS1C1
UBE2L3	PSORS1C2
	RBPJ
	RNF5
	SKIV2L
	SLC22A4
	SLC22A5
	SMAD3
	SNAPC4
	TCF19
	TNFRSF14
	TNFSF15
	TNPO3
	TRAF1
	TRHDE
	TUMF
	UBE2L3
	USP8P1
	ZPBP2

Table 9. Intersections between missSNP harboring genes/ proteins networks and TNF network

Number of intersecting nodes for each pair of genes, followed by gene names (in no specific order)

Genes with missense GWAS AID SNPs		PTPN22	ERAP1	TYK2	CARD9	IL23R	TRAF3IP2	NFKBIE	ATG16L1	RTKN2	TNFAIP3	SH2B3	MST1	LACC1	NOD2	MICA
No of nodes for each gene		**150+**	**50+**	**250+**	**100+**	**130+**	**90**	**50+**	**100+**	**3**	**200+**	**120+**	**100+**	**0**	**150+**	**150+**
TNF		**58**	**17**	**100+**	**50+**	**75**	**34**	**25**	**32**	**0**	**200+**	**23**	**27**	**0**	**80+**	**36**
1500+		INFG CARD9 TNFAIP3 HLA-B IL23R	CARD9 IL6R IL23R HLA-B TNFRSF1A TNFAIP3 INFG ….		IL23R NFKB IL1B TRL2,9 NOD2 TRAF2 STAT3 CARD9 …	IL23R TNFAIP3 CARD9 IL23A INFG HLA-B ….			IL23R NOD2 CARD9 …			JAK2 TNFAIP2 CTLA4 ZAP70 …				HLA KLRK1 …
PTPN22			**8**	**22**	**19**	**45**	**5**	**2**	**18**	**0**	**44**	**30**	**13**	**0**	**31**	**18**
			TNFAIPS IL23R TNIP1 HLA-C FCRL3 HLA-B CARD9 INFG				HLA-C TNFSF13B DEFB4A CD40 TRAF1	TNIP2 TRAF1	NOD2 CARD9 IL23R TNFSF15 …				NOD2 IL23R CARD9 ATG16L1 …			
ERAP1				**5**	**5**	**18**	**6**	**2**	**3**	**0**	**12**	**1**	**2**	**0**	**8**	**5**
				INFG IL6R IL23R IL23A IL28RA	CARD9 IL23R IL23A ERAP1 SNAPC4	SNAPC4 INFG HLA-C TNIP1 IL23R CARD9 ERAP1 TNFAIP3 HLA-B IL23A PHB2 FCRL3 …	TNFRSF1A HLA-C ERAP1 TRAF3IP2 IL28RA BCL2A1 ST20	TNIP1 IL1A	IL23R IL23A CARD9		ERAP1 TNFAIP3 HLA-C TNIP1 BCL2A1 ST20 TNFRSF1A IL23A IL1A PPP1R15A IL23R IL1R2	TNFAIP3	CARD9 IL23R		HLA-C IL23A IL23R CARD9 MRAP TNFAIP3 IFNG IL1A	CD8A HLA-B HLA-C MRAP KLRC2 KLRC3
TYK2					**21**	**50+**	**9**	**7**	**12**	**0**	**29**	**15**	**4**	**0**	**30**	**12**
					PTPN2 TNF IL6 CCR6 IL12B TLR3 FAM92B IL1B STAT3 TRF5 NFKBIA IL17A NFKB2 NFKB1 IL23A IL23R…	STAT3 IL12B IL23R TNF PTPN2 IL1R1 JAK1 IL12A CCR6 JAK2 STAT1 IL6 NFKBIA …	NFKB1 TRAF6 FAM48A IL17A CD40 TNFRSF13C TNFSF13B	TNF NFKB2 RELA NFKB1 CXCL10 FAM48A BTRC	IL1B IL18 CCR6 STAT3 IL12B JAK2 PTPN2 IL23R IL23A …		TRAF6 RELA CTLA4 IL23R TNF NFKB2 NFKB1 STAT4 IL23R IL23A TNFSF13B NFKBIA IL12B MYD88…	LCK TNF IL12A CTLA4…	IL12B IRF5 CXCL10 IL23R		TNF STAT3 IL23R …	NCR1 JAK2 IL15 Il2 STAT3 IL4 IL24 TNF NFKB1 …

CARD9						**31**	**8**	**6**	**39**	**0**	**33**	**5**	**15**	**0**	**49**	**5**
						IL6 PTPN22 NOD1 MST1 ATG16L1 IL1B TLR4 NOD2 IL12B STAT3 SOCS3 ….	TRAF1 TRAF2 IKBKB NFKB1 ERAP1 MAPK8 …	NFKB1 RELB NFKB2 TRAF1 IKBKB TNF	NOD2 MST1 IL23R ATG16L1 CARD9 TNFSF15 …		PTPN22 NFKBIA IL23A TNFIL1B SOCS3 IRF5 ERAP1 NFKB2 RIPK2 IKBKB TLR4 TRAF2 TLR2 TRAF1 NOD2	NOD2 TNF IL18RAP PTPN22 BACE1	NOD2 IL23R CARD9 ATG16L1 IL12B Il18RAP MAPK8 TNFRSF6B IRF5 …		TNF PTPN22 ATG16L1 ….	TNF STAT3 AIRE NFKB1 KLRK1
IL23R							**7**	**4**	**30+**	**0**	**40**	**13**	**14**	**0**	**30+**	**16**
							NFKB1 IL17F DEFB4A ERAP1 IL17A HLA-C FAM48A …	NFKB1 TNF FAM48A TNIP1	PTPN22 STAT3 NOD2 IL23R IRF5 CARD9 …		IL4 IL26 IL6 TNFAIP3 IL13 ERAP1 TNIP1 HLA-C IL23A IL23R NFKBIA SOCS2 NOD2 STAT4 …	PTPN22 JAK2 IL21 MMEL1 TNF CBL TNFAIP3 ADAD1 NOD2 IL12A IL18RAP KIF5A …	DLG5 IL18RAP MST1 IL23R ATG16L1 CARD9 NOD2 TNFSF15 NKX2-3		IL6 NFKB1 TNFSF15 ATG16L1 Il23R ….	HLA-C IL10 IL2 NFKB1 NCR3 ADAD1 IL4 IL15 STAT3 IL24 TNF …
TRAF3IP2								**7**	**1**	**0**	**23**	**0**	**1**	**0**	**8**	**9**
								TRAF1 IKBKB CHUK CD40LG IKBKG FAM48A NFKB1	DEFB4A		TRAF2 IKBKB CHUK TRAF6 MAPK3K7HLA-C AMPK8 ERAP1 NFKB1	none	MAPK8		ERAP1 TRAF1 TRAF2 IL17A, IL17F MAPK8 NFKB1 IKBKB	HLA-DRB1 TRAF3IP3 IL4 TNFRSF25 IL10 HLA-C NFKB1 TNF
NFKBIE									**0**	**0**	**16**	**2**	**3**	**0**	**7**	**2**
									none		CHUK RELB NAF1 ICAM1 NFKB2 IKBKB TNF TRAF1 TNIP1 ILA1 …	AKT1 TNF	PPP2R4 AKT1 CXCL10		TNF IKBKB IKBKG RELA NFKB1 IL1A FAM48A	NFKB1 TNF
ATG16L1										**0**	**19**	**4**	**16**	**0**	**44+**	**3**
											RIPK2 CCNY FADD IL18RAP RIPK1 IL12B TLR9 IL1B IRF5 PTPN22 IL23R TNFSF15 …	JAK2 NOD2 PTPN22 IL18RAP	IRF5 DLG5 IL23R IL12B TNFSF15 CARD9 NKX2-3 NOD2 IL18R		PTPN22 TNFF15 IL23R CASP1 IL18BP ATN1 …	JAK2 STAT3 MYO9B
RTKN2											**0**	**0**	**0**	**0**	**0**	**0**

TNFAIP3													**15**	**9**	**0**	**46**	**9**
													CTLA4 NOD2 PFKFB3 TNF IL21 MMEL1 TNFRSF25 IL21 HLA-DRB1 NAA25 IL18RAP …	NOD2 …		TNF NFKBIA IL1A Il6 TLR10 TNFSF15 APAF1 IRAK1 IL10 RELA HLA-DRB1 IKBKG TRAF6 …	NFKB1 HLA-C IL4 IER3 IL10 TNF TNFRSF25 TRAF3IP#
SH2B3														**4**	**0**	**9**	**9**
														FOXO3 NOD2 IL18RAP AKT1		CTLA4 TNF PTPN22 IL18RAP IL12A HLA-DRAB1 TNFAIP3 NOD2 SH2B3	JAK2 TNF STAT5A TAGAP HLA-DRB1 CD226 ADAD1 DQB1 …
MST1															**0**	**15**	**0**
																IL12B Il23R DLG5 CARD9 ECM1 MARK2 ATG16L1 NKX2-3 …	
LACC1																**0**	**0**
NOD2																	**9**
																	TNF NFKB1 HLA-C LTA HLA-DRB1 …
MICA																	

Note: - missSNP harboring gene/proteins and TNF intersections are described with the number of intersecting nodes for each pair of networks, followed by symbols of some of the genes in the intersections in no specific order

- TNF and IL7R have many common mamabers and TNF among them (not shown in the table)

Table 11. KEGG pathways for the genes harboring missense SNPs

Gene ID	Pathway ID	Pathway name	GWAS related AID
PTPN22	hsa04612	Antigen processing and presentation	RA, CD (T1D, vitiligo)
CARD9	hsa04621	NOD-like receptor signaling pathway	CD, UC, AS
	hsa05152	Tuberculosis	
ERAP1	hsa04612	Antigen processing and presentation	AS, Ps
IL23R	hsa04060	Cytokine-cytokine receptor interaction	CD, AS, Ps, IBD, UC (Behcet syndrome, leprosy, asthma, T1D)
	hsa04630	Jak-STAT signaling pathway	
	hsa05321	Inflammatory bowel disease (IBD)	
TRAF3IP2	hsa04668	TNF signaling pathway	PsA,Ps, CD (T1D, T2D)
	hsa04064	NF-kappa B signaling pathway	
	hsa05160	Hepatitis C	
	hsa05168	Herpes simplex infection	
	hsa05169	Epstein-Barr virus infection	
NFKBIE	hsa04660	T cell receptor signaling pathway	RA
	hsa04662	B cell receptor signaling pathway	
	hsa04064	NF-kappa B signaling pathway	
	hsa04920	Adipocytokine signaling pathway	
	hsa05169	Epstein-Barr virus infection	
MST1	none		CD, UC
ATG16L1	hsa04140	Regulation of autophagy	CD, UC
LACC1	none		CD (leprosy)
TYK2	hsa04380	Osteoclast differentiation	CD, Ps, PsA, UC, IBD, RA (T1D,T2D)
	hsa04630	Jak-STAT signaling pathway	
	hsa05145	Toxoplasmosis	
	hsa05160	Hepatitis C	
	hsa05162	Measles	
	hsa05164	Influenza A	
	hsa05168	Herpes simplex infection	
	hsa05169	Epstein-Barr virus infection	
SH2B3	hsa04722	Neurotrophin signaling pathway	RA (T1D, hypotyroidism)
TNFAIP3	hsa04064	NF-kappa B signaling pathway	RA, PsA, Ps, UC (SLE, celiac disease)
	hsa04621	NOD-like receptor signaling pathway	
	hsa04668	TNF signaling pathway	
	hsa05162	Measles	
	hsa05169	Epstein-Barr virus infection	
NOD2	hsa04621	NOD-like receptor signaling pathway	CD, IBD, Ps

	hsa04668	TNF signaling pathway	
	hsa05131	Shigellosis	
	hsa05152	Tuberculosis	
	hsa05321	Inflammatory bowel disease (IBD)	
IL13	hsa04060	Cytokine-cytokine receptor interaction	Ps (asthma)
	hsa04630	Jak-STAT signaling pathway	
	hsa04664	Fc epsilon RI signaling pathway	
	hsa05162	Measles	
	hsa05310	Asthma	
	hsa05321	Inflammatory bowel disease (IBD)	
IL6R	hsa04060	Cytokine-cytokine receptor interaction	RA (asthma, T1D)
	hsa04151	PI3K-Akt signaling pathway	
	hsa04630	Jak-STAT signaling pathway	
	hsa04640	Hematopoietic cell lineage	
	hsa04932	Non-alcoholic fatty liver disease (NAFLD)	
MICA	none		RA, Ps
RTKN2	none		RA
YDJC	none		RA, CD (celiac disease)
NOTCH4	hsa04330	Notch signaling pathway	Ps (SLE, T1D, T2D, MS, SS, asthma)
FCGR2A	hsa04145	Phagosome	UC (mucokutaneous disease)
	hsa04380	Osteoclast differentiation	
	hsa04611	Platelet activation	
	hsa04666	Fc gamma R-mediated phagocytosis	
	hsa05140	Leishmaniasis	
	hsa05150	Staphylococcus aureus infection	
	hsa05152	Tuberculosis	
	hsa05322	Systemic lupus erythematosus	
GSDMB	none		UC, RA, CD (asthma, T1D)
IL17REL	hsa04060	Cytokine-cytokine receptor interaction	UC, Ps, PsA
IL7R	hsa04060	Cytokine-cytokine receptor interaction	UC (T1D)
	hsa04151	PI3K-Akt signaling pathway	
	hsa04630	Jak-STAT signaling pathway	
	hsa04640	Hematopoietic cell lineage	
	hsa05340	Primary immunodeficiency	

Provided are missSNP harboring genes, labeled with gene symbols, KEGG pathways IDs and association with AIDs

Table 12. KEGG pathways of the genes harboring non-coding SNPs

Gene ID	Pathway ID	Pathway name	GWAS related AID
C2	hsa04610	Complement and coagulation cascades	PS, PsA
	hsa05133	Pertussis	
	hsa05150	Staphylococcus aureus infection	
	hsa05322	Systemic lupus erythematosus	
CCR6	hsa04060	Cytokine-cytokine receptor interaction	RA
	hsa04062	Chemokine signaling pathway	
CD40	hsa04060	Cytokine-cytokine receptor interaction	RA
	hsa04064	NF-kappa B signaling pathway (noncanonical)	
	hsa04514	Cell adhesion molecules (CAMs)	
	hsa04620	Toll-like receptor signaling pathway	
	hsa04672	Intestinal immune network for IgA production	
	hsa05144	Malaria	
	hsa05145	Toxoplasmosis	
	hsa05166	HTLV-I infection	
	hsa05169	Epstein-Barr virus infection	
	hsa05202	Transcriptional misregulation in cancer	
	hsa05310	Asthma	
	hsa05320	Autoimmune thyroid disease	
	hsa05322	Systemic lupus erythematosus	
	hsa05330	Allograft rejection	
	hsa05340	Primary immunodeficiency	
	hsa05416	Viral myocarditis	
	diseases	Immunodef, Hyper IgM syndrom	
CTLA4	hsa04514	Cell adhesion molecules (CAMs)	RA
	hsa04660	T cell receptor signaling pathway	
	hsa05320	Autoimmune thyroid disease	
	hsa05323	Rheumatoid arthritis	
	H00081	Hashimoto's thyroiditis	
	H00082	Graves' disease	
	H00083	Allograft rejection	
	H00408	Type I diabetes mellitus	
CXCR1&2	hsa04060	Cytokine-cytokine receptor interaction	UC
	hsa04062	Chemokine signaling pathway	
	hsa04144	Endocytosis	
	hsa05120	Epithelial cell signaling in Helicobacter pylori infection	

ERAP2	hsa04612	Antigen processing and presentation	CD
IL10	hsa04060	Cytokine-cytokine receptor interaction	UC, CD
	hsa04068	FoxO signaling pathway	
	hsa04630	Jak-STAT signaling pathway	
	hsa04660	T cell receptor signaling pathway	
	hsa04672	Intestinal immune network for IgA production	
	hsa05133	Pertussis	
	hsa05140	Leishmaniasis	
	hsa05142	Chagas disease (American trypanosomiasis)	
	hsa05143	African trypanosomiasis	
	hsa05144	Malaria	
	hsa05145	Toxoplasmosis	
	hsa05146	Amoebiasis	
	hsa05150	Staphylococcus aureus infection	
	hsa05152	Tuberculosis	
	hsa05169	Epstein-Barr virus infection	
	hsa05310	Asthma	
	hsa05320	Autoimmune thyroid disease	
	hsa05321	Inflammatory bowel disease (IBD)	
	hsa05322	Systemic lupus erythematosus	
	hsa05330	Allograft rejection	
IRF1	hsa04917	Prolactin signaling pathway	RA
	hsa05133	Pertussis	
	hsa05160	Hepatitis C (engaged Toll-like R)	
	hsa04620	Toll-like receptor signaling pathway	
IRF5	hsa04620	Toll-like receptor signaling pathway	UC
CXCR1&2	hsa04060	Cytokine-cytokine receptor interaction	UC
	hsa04062	Chemokine signaling pathway	
	hsa04144	Endocytosis	
	hsa05120	Epithelial cell signaling in Helicobacter pylori infection	
LSP1	hsa05152	Tuberculosis	UC
RBPJ	hsa04330	Notch signaling pathway	RA
	hsa05169	Epstein-Barr virus infection	
	hsa05203	Viral carcinogenesis	
TNFSF15	hsa04060	Cytokine-cytokine receptor interaction	CD
	H00286	Crohn's disease	
TNFRSF14	hsa04060	Cytokine-cytokine receptor interaction	UC

	hsa05168	Herpes simplex infection	
TRAF1	hsa04064	NF-kappa B signaling pathway	RA
	hsa04668	TNF signaling pathway	
	hsa05168	Herpes simplex infection	
	hsa05169	Epstein-Barr virus infection	
	hsa05200	Pathways in cancer	
	hsa05202	Transcriptional misregulation in cancer	
	hsa05203	Viral carcinogenesis	
	hsa05222	Small cell lung cancer	
RBPJ	hsa04330	Notch signaling pathway	RA
	hsa05169	Epstein-Barr virus infection	
	hsa05203	Viral carcinogenesis	

Table 13. Classification of AID missSNP KEGG pathways

No	Human KEGG pathways ID	Pathway name	Cathegory of pathways:
1	hsa04060	Cytokine-cytokine receptor interaction	Signaling molecules and interaction; Environmental Information Processing
2	hsa04064	NF-kappa B signaling pathway	Signal transduction; Environmental Information Processing
3	hsa04066	HIF-1 signaling pathway	Signal transduction; Environmental Information Processing
4	hsa04140	Regulation of autophagy	Transport and catabolism; Cellular Processes
5	hsa04145	Phagosome	Transport and catabolism; Cellular Processes
6	hsa04151	PI3K-Akt signaling pathway	Signal transduction; Environmental Information Processing
7	hsa04330	Notch signaling pathway	Signal transduction; Environmental Information Processing
8	hsa04380	Osteoclast differentiation	Development; Organismal Systems
9	hsa04611	Platelet activation	Immune system; Organismal Systems
10	hsa04612	Antigen processing and presentation	Immune system; Organismal Systems
11	hsa04621	NOD-like receptor signaling pathway	Immune system; Organismal Systems
12	hsa04630	Jak-STAT signaling pathway	Signal transduction; Environmental Information Processing
13	hsa04640	Hematopoietic cell lineage	Immune system; Organismal Systems
14	hsa04660	T cell receptor signaling pathway	Immune system; Organismal Systems
15	hsa04662	B cell receptor signaling pathway	Immune system; Organismal Systems
16	hsa04664	Fc epsilon RI signaling pathway	Immune system; Organismal Systems
17	hsa04666	Fc gamma R-mediated phagocytosis	Immune system; Organismal Systems
18	hsa04668	TNF signaling pathway	Signal transduction; Environmental Information Processing
19	hsa04722	Neurotrophin signaling pathway	Nervous system; Organismal Systems
20	hsa04919	Thyroid hormone signaling pathway	Endocrine system; Organismal Systems
21	hsa04920	Adipocytokine signaling pathway	Endocrine system; Organismal Systems
22	hsa04932	Non-alcoholic fatty liver disease (NAFLD)	Endocrine and metabolic diseases; Human Diseases
23	hsa05131	Shigellosis	Infectious diseases: Bacterial
24	hsa05140	Leishmaniasis	Infectious diseases: Parasitic
25	hsa05145	Toxoplasmosis	Infectious diseases: Parasitic
26	hsa05150	Staphylococcus aureus infection	Infectious diseases: Bacterial
27	hsa05152	Tuberculosis	Infectious diseases: Bacterial
28	hsa05160	Hepatitis C	Infectious diseases: Viral
29	hsa05162	Measles	Infectious diseases: Viral
30	hsa05164	Influenza A	Infectious diseases: Viral
31	hsa05168	Herpes simplex infection	Infectious diseases: Viral
32	hsa05169	Epstein-Barr virus infection	Infectious diseases: Viral
33	hsa05206	MicroRNAs in cancer	Cancers: Human Diseases
34	hsa05310	Asthma	Immune diseases Human Diseases
35	hsa05321	Inflammatory bowel disease (IBD)	Immune diseases Human Diseases
36	hsa05340	Primary immunodeficiency	Immune diseases Human Diseases
37	hsa05322	Systemic lupus erythematosus	Immune diseases; Human Diseases
38	H01109	Chronic Mucocutaneous Candidiasis (CMC)	Immune diseases; Human Diseases

Table 14. Classification of KEGG pathways in which ncSNP genes participate

No	Human KEGG pathways ID	Pathway name	Cathegory of pathways:
1	hsa04060	Cytokine-cytokine receptor interaction	Signaling molecules and interaction; Environmental Information Processing
2	hsa04062	Chemokine signaling pathway	Signaling molecules and interaction; Environmental Information Processing
3	hsa04064	NF-kappa B signaling pathway	Signal transduction; Environmental Information Processing
4	hsa04068	FoxO signaling pathway	Signal transduction; Environmental Information Processing
5	hsa04144	Endocytosis	Cellular Processes; Transport and catabolism
6	hsa04330	Notch signaling pathway	Signal transduction; Environmental Information Processing
7	hsa04514	Cell adhesion molecules (CAMs)	Signaling molecules and interaction; Environmental Information Processing
8	hsa04610	Complement and coagulation cascades	Immune system; Organismal Systems
9	hsa04612	Antigen processing and presentation	Immune system; Organismal Systems
10	hsa04620	Toll-like receptor signaling pathway	Immune system; Organismal Systems
11	hsa04630	Jak-STAT signaling pathway	Signal transduction; Environmental Information Processing
12	hsa04660	T cell receptor signaling pathway	Immune system; Organismal Systems
13	hsa04668	TNF signaling pathway	Signal transduction; Environmental Information Processing
14	hsa04672	Intestinal immune network for IgA production	Immune system; Organismal Systems
15	hsa04917	Prolactin signaling pathway	Nervous system; Organismal Systems
16	hsa05120	Epithelial cell signaling in Helicobacter pylori infection	Infectious diseases: Bacterial
17	hsa05133	Pertussis	Infectious diseases: Bacterial
18	hsa05133	Pertussis	Infectious diseases: Bacterial
19	hsa05140	Leishmaniasis	Infectious diseases: Parasitic
20	hsa05142	Chagas disease (American trypanosomiasis)	Infectious diseases: Parasitic
21	hsa05143	African trypanosomiasis	Infectious diseases: Parasitic
22	hsa05144	Malaria	Infectious diseases: Parasitic
23	hsa05145	Toxoplasmosis	Infectious diseases: Parasitic
24	hsa05146	Amoebiasis	Infectious diseases: Parasitic
25	hsa05150	Staphylococcus aureus infection	Infectious diseases: Bacterial
26	hsa05152	Tuberculosis	Infectious diseases: Bacterial
27	hsa05160	Hepatitis C (engaged Toll-like R)	Infectious diseases: Viral
28	hsa05166	HTLV-I infection	Infectious diseases: Viral
29	hsa05168	Herpes simplex infection	Infectious diseases: Viral
30	hsa05169	Epstein-Barr virus infection	Infectious diseases: Viral
31	hsa05200	Pathways in cancer	Cancers: Human Diseases
32	hsa05202	Transcriptional misregulation in cancer	Cancers: Human Diseases
33	hsa05203	Viral carcinogenesis	Cancers: Human Diseases
34	hsa05222	Small cell lung cancer	Cancers: Human Diseases
35	hsa05310	Asthma	Immune diseases; Human Diseases
36	hsa05320	Autoimmune thyroid disease	Immune diseases; Human Diseases
37	hsa05321	Inflammatory bowel disease (IBD)	Immune diseases; Human Diseases
38	hsa05322	Systemic lupus erythematosus	Immune diseases; Human Diseases
39	hsa05323	Rheumatoid arthritis	Immune diseases; Human Diseases
40	hsa05330	Allograft rejection	Immune diseases; Human Diseases
41	hsa05340	Primary immunodeficiency	Immune diseases; Human Diseases
42	hsa05416	Viral myocarditis	Cardiovascular diseases; Human Diseases
43	H00081	Hashimoto's thyroiditis	Immune diseases; Human Diseases
44	H00082	Graves' disease	Immune diseases; Human Diseases
45	H00083	Allograft rejection	Immune diseases; Human Diseases
46	H00286	Crohn's disease	Immune diseases; Human Diseases
47	H00408	Type I diabetes mellitus	Immune diseases; Human Diseases

Table 15. Comparison of five anti-TNF biologic drugs by their known pathways

all	USP drug classification [BR:br08302]				
	Immunological Agents				
	Immune Suppressants				
	Tumor Necrosis Factor (TNF) Blockers				
drug	**ENBREL (Immunex Corporation)**	**REMICADE (Janssen Biotech)**	**HUMIRA (AbbVie)**	**SIMPONI ARIA (Janssen Biotech)**	**CIMZIA (UCB);Certolizumab pegol (genetical recombination) (JAN);**
class	recombinant fusion protein	monoclonal antibody	monoclonal antibody	monoclonal antibody	humanized Fab, pegylated
target	TNF-alpha (hsa:7124)	TNF-alpha (hsa:7124)	TNF-alpha (hsa:7124)	TNF-alpha (hsa:7124)	TNF-alpha (hsa:7124)
	TNF-beta (LT-alpha) (hsa:4049)				
pathways					
	hsa04010 MAPK signaling pathway	hsa04010 MAPK signaling pathway	hsa04010 MAPK signaling pathway	hsa04010 MAPK signaling pathway	hsa04010 MAPK signaling pathway
	hsa04060 Cytokine-cytokine receptor interaction	hsa04060 Cytokine-cytokine receptor interaction	hsa04060 Cytokine-cytokine receptor interaction	hsa04060 Cytokine-cytokine receptor interaction	hsa04060 Cytokine-cytokine receptor interaction
	hsa04210 Apoptosis	hsa04210 Apoptosis	hsa04210 Apoptosis		hsa04210 Apoptosis
	hsa04350 TGF-beta signaling pathway	hsa04350 TGF-beta signaling pathway	hsa04350 TGF-beta signaling pathway	hsa04350 TGF-beta signaling pathway	hsa04350 TGF-beta signaling pathway
	hsa04380 Osteoclast differentiation	hsa04380 Osteoclast differentiation			
	hsa04612 Antigen processing and presentation	hsa04612 Antigen processing and presentation	hsa04612 Antigen processing and presentation		hsa04612 Antigen processing and presentation
	hsa04920 Adipocytokine signaling pathway	hsa04920 Adipocytokine signaling pathway		hsa04920 Adipocytokine signaling pathway	hsa04920 Adipocytokine signaling pathway
	hsa05323 Rheumatoid arthritis	hsa05323 Rheumatoid arthritis	hsa05323 Rheumatoid arthritis	hsa05323 Rheumatoid arthritis	hsa05323 Rheumatoid arthritis
			hsa04650 Natural killer cell mediated cytotoxicity		hsa04650 Natural killer cell mediated cytotoxicity
				hsa04664 Fc epsilon RI signaling pathway	
				hsa05310 Asthma	
					hsa04620 Toll-like receptor signaling pathway
activity	To decrease signs and symptoms of rheumatoid arthritis [DS:H00287 H00288 H00630]	not defined	Treatment of rheumatoid arthritis and other chronic inflammatory disease	Treatment of inflammatory disorders, rheumatoid arthritis, uveitis, asthma and Crohn's disease	Treatment of rheumatoid arthritis and inflammatory bowel disease, specifically Crohn's disease

USP classification; name of biologics and manufacturer; class and targets of biologics; detected pathways; action of each biologics.

Table 16. TNF human pathways from KEGG

Hits 45 from KEGG database

KEGG PATHWAY that have TNF as an participant		Is TNF a product or a signal in the pathway?
hsa04010	MAPK signaling pathway - Homo sapiens (human)	signal (only partially, no Jak-STAT))
hsa04060	Cytokine-cytokine receptor interaction - Homo sapiens (human)	collection of proteins
hsa04064	NF-kappa B signaling pathway - Homo sapiens (human)	signal
hsa04150	mTOR signaling pathway - Homo sapiens (human)	only probable signal
hsa04210	Apoptosis - Homo sapiens (human)	signal
hsa04350	TGF-beta signaling pathway - Homo sapiens (human)	signal
hsa04380	Osteoclast differentiation - Homo sapiens (human)	signal
hsa04612	Antigen processing and presentation - Homo sapiens (human)	signal
hsa04620	Toll-like receptor signaling pathway - Homo sapiens (human)	product
hsa04621	NOD-like receptor signaling pathway - Homo sapiens (human)	product
hsa04622	RIG-I-like receptor signaling pathway - Homo sapiens (human)	product
hsa04640	Hematopoietic cell lineage - Homo sapiens (human)	signal
hsa04650	Natural killer cell mediated cytotoxicity - Homo sapiens (human)	product
hsa04660	T cell receptor signaling pathway - Homo sapiens (human)	product
hsa04664	Fc epsilon RI signaling pathway - Homo sapiens (human)	product
hsa04668	TNF signaling pathway - Homo sapiens (human)	signal
hsa04920	Adipocytokine signaling pathway - Homo sapiens (human)	signal (other half is Jak-STAT dipendant)
hsa04930	Type II diabetes mellitus - Homo sapiens (human)	signal
hsa04932	Non-alcoholic fatty liver disease (NAFLD) - Homo sapiens (human)	signal and product
hsa04940	Type I diabetes mellitus - Homo sapiens (human)	signal and product
hsa05010	Alzheimer's disease - Homo sapiens (human)	na
hsa05014	Amyotrophic lateral sclerosis (ALS) - Homo sapiens (human)	na
hsa05133	Pertussis - Homo sapiens (human)	na
hsa05134	Legionellosis - Homo sapiens (human)	na
hsa05140	Leishmaniasis - Homo sapiens (human)	na
hsa05142	Chagas disease (American trypanosomiasis) - Homo sapiens (human)	na
hsa05143	African trypanosomiasis - Homo sapiens (human)	na
hsa05144	Malaria - Homo sapiens (human)	na
hsa05145	Toxoplasmosis - Homo sapiens (human)	na
hsa05146	Amoebiasis - Homo sapiens (human)	na
hsa05152	Tuberculosis - Homo sapiens (human)	signal
hsa05160	Hepatitis C - Homo sapiens (human)	signal
hsa05161	Hepatitis B - Homo sapiens (human)	product
hsa05164	Influenza A - Homo sapiens (human)	product
hsa05166	HTLV-I infection - Homo sapiens (human)	signal
hsa05168	Herpes simplex infection - Homo sapiens (human)	product
hsa05205	Proteoglycans in cancer - Homo sapiens (human)	na
hsa05310	Asthma - Homo sapiens (human)	na
hsa05321	Inflammatory bowel disease (IBD) - Homo sapiens (human)	na
hsa05322	Systemic lupus erythematosus - Homo sapiens (human)	na
hsa05323	Rheumatoid arthritis - Homo sapiens (human)	na
hsa05330	Allograft rejection - Homo sapiens (human)	na
hsa05332	Graft-versus-host disease - Homo sapiens (human)	na
hsa05410	Hypertrophic cardiomyopathy (HCM) - Homo sapiens (human)	na
hsa05414	Dilated cardiomyopathy - Homo sapiens (human)	na

KEGG pathway ID and name; position of TNF (as a signal or a product) in a pathway where data are available.
na: does not apply

Table 17.
Relationship between AIDSNP pathways and TNF pathways

Pathways common to all AID SNP pathways and TNF pathways

No.	KEGG pathway ID	KEGG pathway name
1	hsa04010	MAPK signaling pathway
2	hsa04060	Cytokine-cytokine receptor interaction
3	hsa04064	NF-kappa B signaling pathway
4	hsa04210	Apoptosis (survival mode by NFkB pathway)
5	hsa04350	TGF-beta signaling pathway
6	hsa04380	Osteoclast differentiation
7	hsa04612	Antigen processing and presentation
8	hsa04620	Toll-like receptor signaling pathway
9	hsa04621	NOD-like receptor signaling pathway
10	hsa04640	Hematopoietic cell lineage
11	hsa04660	T cell receptor signaling pathway
12	hsa04664	Fc epsilon RI signaling pathway
13	hsa04668	TNF signaling pathway
14	hsa04920	Adipocytokine signaling pathway

Pathways common to AID SNP pathways but not TNF pathways

No.	KEGG pathway ID	KEGG pathway name
1	hsa04120	Ubiquitin mediated proteolysis
2	hsa04630	Jak-STAT signaling pathway
3	hsa04330	Notch signaling pathway
4	hsa04672	Intestinal immune network for IgA production
5	hsa04662	B cell receptor signaling pathway

Note:

	pathways common to ncSNP and TNF pathways
	pathways common to missSNP, ncSNP and TNF pathways
	pathways common to missSNP and TNF pathways
	missSNP pathways or ncSNP pathways, but not TNF
	pathways common to missSNP pathways and ncSNP pathways, but not TNF

Table 18. STRING pathways enrichment data for AID GWAS SNP genesets

KEGG ID	Term name of the KEGG pathway	Number Of Genes	p-value	p-value_fdr	p-value_bonferroni
missSNPs set					
hsa04630	Jak-STAT signaling pathway	5	3.25E-07	7.70E-05	7.70E-05
hsa04621	NOD-like receptor signaling pathway	3	2.17E-05	2.57E-03	5.14E-03
hsa04060	Cytokine-cytokine receptor interaction	4	1.08E-04	8.51E-03	2.55E-02
hsa04380	Osteoclast differentiation	3	2.16E-04	1.02E-02	5.12E-02
hsa05162	Measles	3	2.67E-04	1.58E-02	6.33E-02
hsa05152	Tuberculosis	3	6.16E-04	2.92E-02	1.46E-01
ncSNPs set					
hsa04060	Cytokine-cytokine receptor interaction	4	1.54E-04	3.64E-02	3.64E-02
all SNPs set					
hsa04060	Cytokine-cytokine receptor interaction	8	4.22E-08	1.00E-05	1.00E-05
hsa04630	Jak-STAT signaling pathway	6	5.83E-07	6.91E-05	1.38E-04
hsa04621	NOD-like receptor signaling pathway	4	4.73E-06	3.74E-04	1.12E-03
hsa05162	Measles	4	1.31E-04	7.79E-03	3.11E-02
hsa05120	Epithelial cell signaling in H. pylori infect	3	2.99E-04	1.33E-02	7.07E-02
hsa05140	Leishmaniasis	3	3.87E-04	1.33E-02	9.18E-02
hsa05152	Tuberculosis	4	3.92E-04	1.33E-02	9.29E-02
hsa04062	Chemokine signaling pathway	4	4.74E-04	1.41E-02	1.12E-01
hsa04660	T cell receptor signaling pathway	3	1.21E-03	3.19E-02	2.87E-01
hsa05310	Asthma	2	1.64E-03	3.88E-02	3.88E-01
hsa04380	Osteoclast differentiation	3	1.82E-03	3.92E-02	4.31E-01
hsa05160	Hepatitis C	3	2.19E-03	4.07E-02	5.18E-01
hsa05145	Toxoplasmosis	3	2.23E-03	4.07E-02	5.29E-01

KEGG pathways IDs and names are given; number of genes in a gene set for each pathway; p values, after FDR and Bonferroni corrections.

yellow background signs statistical significance for FDR and Bonferron

Table 19. Enriched KEGG pathway-based set for missSNP geneset by ConsensusPathDB

uploaded list:	22
mapped entities:	20
enriched pathway-based sets:	23

18 genes (90.0%) from the input list are present in at least one pathway

pathway name	set size	candidates contained	p-value	q-value	pathway source
NOD-like receptor signaling pathway - Homo sapiens (human)	57	5 (8.8%)	2.82E-08	9.30E-07	KEGG
Inflammatory bowel disease (IBD) - Homo sapiens (human)	67	4 (6.3%)	3.06E-06	3.61E-05	KEGG
Jak-STAT signaling pathway - Homo sapiens (human)	156	5 (3.4%)	3.28E-06	3.61E-05	KEGG
TNF signaling pathway - Homo sapiens (human)	110	4 (3.7%)	2.72E-05	0.000224	KEGG
Osteoclast differentiation - Homo sapiens (human)	131	4 (3.3%)	4.10E-05	0.000229	KEGG
Cytokine-cytokine receptor interaction - Homo sapiens (human)	265	5 (2.0%)	4.16E-05	0.000229	KEGG
Measles - Homo sapiens (human)	134	4 (3.1%)	5.43E-05	0.000256	KEGG
Tuberculosis - Homo sapiens (human)	179	4 (2.3%)	0.000165	0.00068	KEGG
Leishmaniasis - Homo sapiens (human)	74	3 (4.2%)	0.000217	0.000797	KEGG
Hematopoietic cell lineage - Homo sapiens (human)	87	3 (3.7%)	0.000319	0.00105	KEGG
NF-kappa B signaling pathway - Homo sapiens (human)	91	3 (3.4%)	0.000393	0.00118	KEGG
Toxoplasmosis - Homo sapiens (human)	120	3 (2.6%)	0.00086	0.00237	KEGG
Asthma - Homo sapiens (human)	32	2 (6.7%)	0.00112	0.00283	KEGG
Hepatitis C - Homo sapiens (human)	133	3 (2.3%)	0.0012	0.00283	KEGG
Influenza A - Homo sapiens (human)	177	3 (1.8%)	0.0026	0.00572	KEGG
Herpes simplex infection - Homo sapiens (human)	186	3 (1.7%)	0.00306	0.00632	KEGG
Legionellosis - Homo sapiens (human)	55	2 (3.7%)	0.00358	0.00696	KEGG
Epstein-Barr virus infection - Homo sapiens (human)	202	3 (1.5%)	0.00407	0.00746	KEGG
Shigellosis - Homo sapiens (human)	61	2 (3.3%)	0.00455	0.00791	KEGG
Fc epsilon RI signaling pathway - Homo sapiens (human)	70	2 (2.9%)	0.00563	0.00868	KEGG
RIG-I-like receptor signaling pathway - Homo sapiens (human)	70	2 (2.9%)	0.00579	0.00868	KEGG
Adipocytokine signaling pathway - Homo sapiens (human)	70	2 (2.9%)	0.00579	0.00868	KEGG
Apoptosis - Homo sapiens (human)	86	2 (2.5%)	0.0079	0.0113	KEGG

Number of queried genes entered for enrichment analyses and number of genes mapped in pathways; number of enriched pathway sets; pathways names with the number of genes in each pathway and percentage of queried genes for each pathway along with number of genes in a gene set for each pathway; p values, after fdr and Bonferroni corrections.

Table 20. Enriched KEGG pathway-based sets for allSNP set by ConsensusPathDB

uploaded list: 56
mapped entities: 54
enriched pathway-based sets: 27

39 genes (72.2%) from the input list are present in at least one pathway.

pathway name	set size	candidates contained	p-value	q-value	pathway source
Cytokine-cytokine receptor interaction - Homo sapiens (human)	265	11 (4.4%)	7.32E-10	3.58E-08	KEGG
Epstein-Barr virus infection - Homo sapiens (human)	202	8 (4.0%)	4.27E-07	1.05E-05	KEGG
Inflammatory bowel disease (IBD) - Homo sapiens (human)	67	5 (7.9%)	2.88E-06	4.70E-05	KEGG
Jak-STAT signaling pathway - Homo sapiens (human)	156	6 (4.1%)	1.31E-05	0.000161	KEGG
Autoimmune thyroid disease - Homo sapiens (human)	54	4 (7.7%)	3.50E-05	0.000343	KEGG
NOD-like receptor signaling pathway - Homo sapiens (human)	57	4 (7.0%)	5.05E-05	0.000412	KEGG
Asthma - Homo sapiens (human)	32	3 (10.0%)	0.000166	0.00117	KEGG
NF-kappa B signaling pathway - Homo sapiens (human)	91	4 (4.5%)	0.000275	0.00168	KEGG
Allograft rejection - Homo sapiens (human)	39	3 (8.1%)	0.000313	0.0017	KEGG
Herpes simplex infection - Homo sapiens (human)	186	5 (2.8%)	0.000436	0.00214	KEGG
TNF signaling pathway - Homo sapiens (human)	110	4 (3.7%)	0.000621	0.00277	KEGG
Toxoplasmosis - Homo sapiens (human)	120	4 (3.5%)	0.00076	0.0031	KEGG
Staphylococcus aureus infection - Homo sapiens (human)	57	3 (5.7%)	0.000907	0.00342	KEGG
Measles - Homo sapiens (human)	134	4 (3.1%)	0.0012	0.0042	KEGG
Systemic lupus erythematosus - Homo sapiens (human)	136	4 (3.0%)	0.00131	0.00427	KEGG
Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human)	68	3 (4.5%)	0.00179	0.00548	KEGG
Leishmaniasis - Homo sapiens (human)	74	3 (4.2%)	0.0022	0.00635	KEGG
Pertussis - Homo sapiens (human)	75	3 (4.0%)	0.00247	0.00673	KEGG
Tuberculosis - Homo sapiens (human)	179	4 (2.3%)	0.00341	0.00879	KEGG
Chemokine signaling pathway - Homo sapiens (human)	189	4 (2.2%)	0.00417	0.0102	KEGG
T cell receptor signaling pathway - Homo sapiens (human)	104	3 (3.0%)	0.00572	0.013	KEGG
Chagas disease (American trypanosomiasis) - Homo sapiens (human)	104	3 (2.9%)	0.00604	0.013	KEGG
Viral carcinogenesis - Homo sapiens (human)	206	4 (2.0%)	0.00622	0.013	KEGG
Toll-like receptor signaling pathway - Homo sapiens (human)	106	3 (2.9%)	0.00637	0.013	KEGG
Endocytosis - Homo sapiens (human)	213	4 (1.9%)	0.00677	0.0133	KEGG
Primary immunodeficiency - Homo sapiens (human)	36	2 (5.7%)	0.00705	0.0133	KEGG
Osteoclast differentiation - Homo sapiens (human)	131	3 (2.5%)	0.00941	0.0171	KEGG

Number of queried genes entered for enrichment analyses and number of genes mapped in pathways; number of enriched pathway sets; pathways names with the number of genes in each pathway and percentage of queried genes for each pathway along with number of genes in a gene set for each pathway; p values, after fdr and Bonferroni corrections.

Table 21. GO term BP, MF and CC annotation enrichment for AID SNP sets

(yellow highlited are significant terms based on p values for Bonferroni or FDR)

missSNP set **BP terms**

GO_id		No of genes	p-value	p-value_fdr	p-value_bonferroni
GO:0002376	immune system process	12	1.73E-08	2.18E-04	2.18E-04
GO:0009617	response to bacterium	7	4.61E-08	2.91E-04	5.81E-04
GO:0002684	positive regulation of immune system process	8	1.26E-07	5.28E-04	1.58E-03
GO:0001818	negative regulation of cytokine production	5	2.24E-07	7.07E-04	2.83E-03
GO:0002682	regulation of immune system process	9	3.05E-07	7.70E-04	3.85E-03
GO:0006955	immune response	9	8.23E-07	1.33E-03	1.04E-02
GO:0051707	response to other organism	7	8.43E-07	1.33E-03	1.06E-02
GO:0043207	response to external biotic stimulus	7	8.43E-07	1.33E-03	1.06E-02
GO:0009607	response to biotic stimulus	7	1.17E-06	1.64E-03	1.48E-02
GO:0045088	regulation of innate immune response	5	2.23E-06	2.81E-03	2.81E-02
GO:0070423	nucleotide-binding oligomerization domain containing signaling pa	3	2.61E-06	2.87E-03	3.29E-02
GO:0034136	negative regulation of toll-like receptor 2 signaling pathway	2	2.73E-06	2.87E-03	3.44E-02
GO:0050776	regulation of immune response	7	3.92E-06	3.81E-03	4.95E-02
GO:0002753	cytoplasmic pattern recognition receptor signaling pathway	3	4.75E-06	3.97E-03	5.98E-02
GO:0002822	regulation of adaptive immune response based on somatic recom	4	4.87E-06	3.97E-03	6.14E-02
GO:0002697	regulation of immune effector process	5	5.03E-06	3.97E-03	6.35E-02
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	4	6.12E-06	4.42E-03	7.72E-02
GO:0002819	regulation of adaptive immune response	4	6.32E-06	4.42E-03	7.96E-02

ncSNP set

GO_id		No of genes	p-value	p-value_fdr	p-value_bonferroni
GO:0002520	immune system development	7	3.13E-07	2.48E-03	3.94E-03
GO:0002407	dendritic cell chemotaxis	3	4.80E-07	2.48E-03	6.05E-03
GO:0036336	dendritic cell migration	3	5.90E-07	2.48E-03	7.44E-03
GO:0038112	interleukin-8-mediated signaling pathway	2	1.09E-06	2.59E-03	1.37E-02
GO:0001775	cell activation	7	1.16E-06	2.59E-03	1.46E-02
GO:0045321	leukocyte activation	6	1.23E-06	2.59E-03	1.55E-02
GO:0030097	hemopoiesis	6	2.60E-06	4.69E-03	3.28E-02
GO:0048534	hematopoietic or lymphoid organ development	6	4.58E-06	6.51E-03	5.77E-02
GO:0002521	leukocyte differentiation	5	4.65E-06	6.51E-03	5.86E-02
GO:0007221	positive regulation of transcription of Notch receptor target	2	6.52E-06	8.22E-03	8.22E-02
GO:0046649	lymphocyte activation	5	9.86E-06	1.13E-02	1.24E-01
GO:0032655	regulation of interleukin-12 production	3	1.19E-05	1.25E-02	1.50E-01
GO:0043117	positive regulation of vascular permeability	2	3.04E-05	2.79E-02	3.83E-01
GO:0030098	lymphocyte differentiation	4	3.10E-05	2.79E-02	3.90E-01
GO:0009611	response to wounding	7	3.71E-05	3.12E-02	4.68E-01
GO:0043112	receptor metabolic process	3	4.45E-05	3.50E-02	5.60E-01
GO:0071350	cellular response to interleukin-15	2	4.87E-05	3.61E-02	6.14E-01
GO:0050900	leukocyte migration	4	6.74E-05	4.71E-02	8.50E-01
GO:0030183	B cell differentiation	3	7.10E-05	4.71E-02	8.95E-01

allSNP set

GO_id		No of genes	p-value	p-value_fdr	p-value_bonferroni
GO:0006955	immune response	17	4.64E-11	5.85E-07	5.85E-07
GO:0002520	immune system development	11	9.88E-10	6.22E-06	1.24E-05
GO:0009617	response to bacterium	10	2.76E-09	8.91E-06	3.48E-05
GO:0002376	immune system process	18	2.83E-09	8.91E-06	3.56E-05
GO:0032495	response to muramyl dipeptide	4	4.54E-09	1.14E-05	5.72E-05
GO:0043207	response to external biotic stimulus	11	1.28E-08	2.26E-05	1.62E-04
GO:0051707	response to other organism	11	1.28E-08	2.26E-05	1.62E-04
GO:0009605	response to external stimulus	16	1.44E-08	2.26E-05	1.81E-04
GO:0032655	regulation of interleukin-12 production	5	2.04E-08	2.44E-05	2.57E-04
GO:0045321	leukocyte activation	9	2.06E-08	2.44E-05	2.60E-04

GO:0009607	response to biotic stimulus	11	2.13E-08	2.44E-05	2.68E-04
GO:0030097	hemopoiesis	9	6.24E-08	6.56E-05	7.87E-04
GO:0002682	regulation of immune system process	13	8.95E-08	8.68E-05	1.13E-03
GO:0001775	cell activation	10	9.77E-08	8.80E-05	1.23E-03
GO:0098542	defense response to other organism	8	1.27E-07	1.07E-04	1.60E-03
GO:0048534	hematopoietic or lymphoid organ development	9	1.44E-07	1.13E-04	1.81E-03
GO:0050777	negative regulation of immune response	5	1.77E-07	1.23E-04	2.23E-03
GO:0009611	response to wounding	12	1.85E-07	1.23E-04	2.33E-03
GO:0002819	regulation of adaptive immune response	6	1.85E-07	1.23E-04	2.34E-03
GO:0031347	regulation of defense response	9	2.54E-07	1.60E-04	3.21E-03
GO:0070423	nucleotide-binding oligomerization domain containing signaling pa	4	2.75E-07	1.65E-04	3.46E-03
GO:0002237	response to molecule of bacterial origin	7	3.16E-07	1.81E-04	3.98E-03
GO:0001819	positive regulation of cytokine production	7	4.59E-07	2.52E-04	5.79E-03
GO:0002684	positive regulation of immune system process	10	6.07E-07	3.11E-04	7.65E-03
GO:0002753	cytoplasmic pattern recognition receptor signaling pathway	4	6.17E-07	3.11E-04	7.78E-03
GO:0050864	regulation of B cell activation	5	8.77E-07	4.13E-04	1.11E-02
GO:0071350	cellular response to interleukin-15	3	8.86E-07	4.13E-04	1.12E-02
GO:0070663	regulation of leukocyte proliferation	6	9.40E-07	4.23E-04	1.18E-02
GO:0042742	defense response to bacterium	6	1.01E-06	4.38E-04	1.27E-02
GO:0046649	lymphocyte activation	7	1.26E-06	5.29E-04	1.59E-02
GO:0050776	regulation of immune response	10	1.31E-06	5.32E-04	1.65E-02
GO:0001817	regulation of cytokine production	8	1.64E-06	6.46E-04	2.07E-02
GO:0002252	immune effector process	8	1.82E-06	6.96E-04	2.30E-02
GO:0035872	nucleotide-binding domain, leucine rich repeat containing recepto	4	1.95E-06	7.23E-04	2.46E-02
GO:0070672	response to interleukin-15	3	2.10E-06	7.57E-04	2.65E-02
GO:0051249	regulation of lymphocyte activation	7	2.38E-06	8.32E-04	3.00E-02
GO:0042127	regulation of cell proliferation	12	2.94E-06	1.00E-03	3.71E-02
GO:0002407	dendritic cell chemotaxis	3	3.33E-06	1.11E-03	4.20E-02
GO:0002822	regulation of adaptive immune response based on somatic recom	5	3.75E-06	1.19E-03	4.72E-02
GO:0043331	response to dsRNA	4	3.79E-06	1.19E-03	4.78E-02
GO:0038112	interleukin-8-mediated signaling pathway	2	3.90E-06	1.20E-03	4.92E-02
GO:0036336	dendritic cell migration	3	4.10E-06	1.23E-03	5.17E-02
GO:0032496	response to lipopolysaccharide	6	4.33E-06	1.27E-03	5.45E-02
GO:0006954	inflammatory response	7	5.42E-06	1.55E-03	6.83E-02
GO:0045088	regulation of innate immune response	6	5.61E-06	1.57E-03	7.07E-02
GO:0002521	leukocyte differentiation	6	7.72E-06	2.12E-03	9.74E-02
GO:0006950	response to stress	17	7.99E-06	2.14E-03	1.01E-01
GO:0001818	negative regulation of cytokine production	5	9.26E-06	2.43E-03	1.17E-01
GO:0034136	negative regulation of toll-like receptor 2 signaling pathway	2	1.17E-05	3.01E-03	1.47E-01
GO:0002697	regulation of immune effector process	6	1.45E-05	3.62E-03	1.83E-01
GO:0032735	positive regulation of interleukin-12 production	3	1.46E-05	3.62E-03	1.85E-01
GO:0006952	defense response	11	1.54E-05	3.72E-03	1.94E-01
GO:0050670	regulation of lymphocyte proliferation	5	1.65E-05	3.93E-03	2.08E-01
GO:0032944	regulation of mononuclear cell proliferation	5	1.75E-05	4.08E-03	2.21E-01
GO:0032675	regulation of interleukin-6 production	4	1.78E-05	4.08E-03	2.24E-01
GO:0007221	positive regulation of transcription of Notch receptor target	2	2.34E-05	5.08E-03	2.94E-01
GO:0002677	negative regulation of chronic inflammatory response	2	2.34E-05	5.08E-03	2.94E-01
GO:0030098	lymphocyte differentiation	5	2.34E-05	5.08E-03	2.95E-01
GO:0030595	leukocyte chemotaxis	4	2.39E-05	5.10E-03	3.01E-01
GO:0050869	negative regulation of B cell activation	3	2.63E-05	5.52E-03	3.31E-01
GO:0032720	negative regulation of tumor necrosis factor production	3	2.91E-05	6.02E-03	3.67E-01
GO:0080134	regulation of response to stress	9	3.20E-05	6.39E-03	4.03E-01
GO:0045087	innate immune response	9	3.20E-05	6.39E-03	4.03E-01
GO:0042110	T cell activation	5	3.67E-05	7.23E-03	4.63E-01
GO:0070431	nucleotide-binding oligomerization domain containing 2 signaling	2	3.89E-05	7.43E-03	4.90E-01
GO:0034135	regulation of toll-like receptor 2 signaling pathway	2	3.89E-05	7.43E-03	4.90E-01
GO:0032479	regulation of type I interferon production	4	4.04E-05	7.60E-03	5.09E-01
GO:0014070	response to organic cyclic compound	8	4.26E-05	7.90E-03	5.37E-01
GO:0071219	cellular response to molecule of bacterial origin	4	5.74E-05	1.04E-02	7.23E-01
GO:2000026	regulation of multicellular organismal development	10	5.76E-05	1.04E-02	7.27E-01
GO:0043330	response to exogenous dsRNA	3	5.99E-05	1.05E-02	7.55E-01

GO:0050900	leukocyte migration	5	6.06E-05	1.05E-02	7.64E-01
GO:0002757	immune response-activating signal transduction	6	6.09E-05	1.05E-02	7.67E-01
GO:1901698	response to nitrogen compound	8	6.20E-05	1.05E-02	7.81E-01
GO:0048584	positive regulation of response to stimulus	11	6.27E-05	1.05E-02	7.91E-01
GO:0002694	regulation of leukocyte activation	6	6.49E-05	1.08E-02	8.18E-01
GO:0050920	regulation of chemotaxis	4	6.88E-05	1.11E-02	8.67E-01
GO:0060326	cell chemotaxis	4	6.88E-05	1.11E-02	8.67E-01
GO:0032663	regulation of interleukin-2 production	3	6.99E-05	1.12E-02	8.81E-01
GO:0002706	regulation of lymphocyte mediated immunity	4	7.38E-05	1.16E-02	9.30E-01
GO:0042531	positive regulation of tyrosine phosphorylation of STAT protein	3	8.10E-05	1.22E-02	1.00E+00
GO:0002377	immunoglobulin production	3	8.10E-05	1.22E-02	1.00E+00
GO:0010536	positive regulation of activation of Janus kinase activity	2	8.14E-05	1.22E-02	1.00E+00
GO:0032817	regulation of natural killer cell proliferation	2	8.14E-05	1.22E-02	1.00E+00
GO:0051251	positive regulation of lymphocyte activation	5	8.57E-05	1.27E-02	1.00E+00
GO:0051240	positive regulation of multicellular organismal process	7	9.02E-05	1.30E-02	1.00E+00
GO:0042113	B cell activation	4	9.04E-05	1.30E-02	1.00E+00
GO:0048583	regulation of response to stimulus	15	9.11E-05	1.30E-02	1.00E+00
GO:0050863	regulation of T cell activation	5	9.90E-05	1.39E-02	1.00E+00
GO:0050865	regulation of cell activation	6	9.92E-05	1.39E-02	1.00E+00
GO:0071216	cellular response to biotic stimulus	4	1.03E-04	1.42E-02	1.00E+00
GO:0050793	regulation of developmental process	11	1.08E-04	1.45E-02	1.00E+00
GO:0043117	positive regulation of vascular permeability	2	1.08E-04	1.45E-02	1.00E+00
GO:0010533	regulation of activation of Janus kinase activity	2	1.08E-04	1.45E-02	1.00E+00
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	4	1.13E-04	1.50E-02	1.00E+00
GO:0042509	regulation of tyrosine phosphorylation of STAT protein	3	1.21E-04	1.59E-02	1.00E+00
GO:0002696	positive regulation of leukocyte activation	5	1.25E-04	1.63E-02	1.00E+00
GO:0032703	negative regulation of interleukin-2 production	2	1.39E-04	1.77E-02	1.00E+00
GO:0002676	regulation of chronic inflammatory response	2	1.39E-04	1.77E-02	1.00E+00
GO:0046427	positive regulation of JAK-STAT cascade	3	1.45E-04	1.80E-02	1.00E+00
GO:0050830	defense response to Gram-positive bacterium	3	1.45E-04	1.80E-02	1.00E+00
GO:0050867	positive regulation of cell activation	5	1.46E-04	1.80E-02	1.00E+00
GO:0002703	regulation of leukocyte mediated immunity	4	1.52E-04	1.86E-02	1.00E+00
GO:0030888	regulation of B cell proliferation	3	1.54E-04	1.86E-02	1.00E+00
GO:0051607	defense response to virus	4	1.61E-04	1.93E-02	1.00E+00
GO:0032494	response to peptidoglycan	2	1.74E-04	2.05E-02	1.00E+00
GO:0032740	positive regulation of interleukin-17 production	2	1.74E-04	2.05E-02	1.00E+00
GO:0070664	negative regulation of leukocyte proliferation	3	1.92E-04	2.24E-02	1.00E+00
GO:0002710	negative regulation of T cell mediated immunity	2	2.12E-04	2.40E-02	1.00E+00
GO:0032695	negative regulation of interleukin-12 production	2	2.12E-04	2.40E-02	1.00E+00
GO:0002862	negative regulation of inflammatory response to antigenic stimulu	2	2.12E-04	2.40E-02	1.00E+00
GO:0071345	cellular response to cytokine stimulus	6	2.13E-04	2.40E-02	1.00E+00
GO:0033993	response to lipid	7	2.27E-04	2.53E-02	1.00E+00
GO:0035556	intracellular signal transduction	10	2.44E-04	2.69E-02	1.00E+00
GO:0032088	negative regulation of NF-kappaB transcription factor activity	3	2.47E-04	2.69E-02	1.00E+00
GO:0002688	regulation of leukocyte chemotaxis	3	2.47E-04	2.69E-02	1.00E+00
GO:0050795	regulation of behavior	4	2.58E-04	2.78E-02	1.00E+00
GO:0010243	response to organonitrogen compound	7	2.67E-04	2.84E-02	1.00E+00
GO:0002764	immune response-regulating signaling pathway	6	2.68E-04	2.84E-02	1.00E+00
GO:0043112	receptor metabolic process	3	2.99E-04	3.05E-02	1.00E+00
GO:0042534	regulation of tumor necrosis factor biosynthetic process	2	3.00E-04	3.05E-02	1.00E+00
GO:0052200	response to host defenses	2	3.00E-04	3.05E-02	1.00E+00
GO:0052173	response to defenses of other organism involved in symbiotic inter	2	3.00E-04	3.05E-02	1.00E+00
GO:0075136	response to host	2	3.00E-04	3.05E-02	1.00E+00
GO:0071260	cellular response to mechanical stimulus	3	3.12E-04	3.15E-02	1.00E+00
GO:2000106	regulation of leukocyte apoptotic process	3	3.26E-04	3.27E-02	1.00E+00
GO:0002709	regulation of T cell mediated immunity	3	3.41E-04	3.39E-02	1.00E+00
GO:0032647	regulation of interferon-alpha production	2	3.50E-04	3.43E-02	1.00E+00
GO:0030522	intracellular receptor signaling pathway	4	3.51E-04	3.43E-02	1.00E+00
GO:0050730	regulation of peptidyl-tyrosine phosphorylation	4	3.59E-04	3.48E-02	1.00E+00
GO:0032103	positive regulation of response to external stimulus	4	3.83E-04	3.67E-02	1.00E+00
GO:0048660	regulation of smooth muscle cell proliferation	3	3.87E-04	3.67E-02	1.00E+00

GO:0002700	regulation of production of molecular mediator of immune respon	3	3.87E-04	3.67E-02	1.00E+00
GO:0050870	positive regulation of T cell activation	4	4.00E-04	3.69E-02	1.00E+00
GO:0030889	negative regulation of B cell proliferation	2	4.03E-04	3.69E-02	1.00E+00
GO:0033033	negative regulation of myeloid cell apoptotic process	2	4.03E-04	3.69E-02	1.00E+00
GO:0048535	lymph node development	2	4.03E-04	3.69E-02	1.00E+00
GO:0046425	regulation of JAK-STAT cascade	3	4.04E-04	3.69E-02	1.00E+00
GO:0050778	positive regulation of immune response	6	4.12E-04	3.74E-02	1.00E+00
GO:0032680	regulation of tumor necrosis factor production	3	4.20E-04	3.78E-02	1.00E+00
GO:0032649	regulation of interferon-gamma production	3	4.55E-04	4.00E-02	1.00E+00
GO:0032660	regulation of interleukin-17 production	2	4.60E-04	4.00E-02	1.00E+00
GO:0002823	negative regulation of adaptive immune response based on somati	2	4.60E-04	4.00E-02	1.00E+00
GO:0030183	B cell differentiation	3	4.73E-04	4.09E-02	1.00E+00
GO:0002707	negative regulation of lymphocyte mediated immunity	2	5.21E-04	4.34E-02	1.00E+00
GO:0044130	negative regulation of growth of symbiont in host	2	5.21E-04	4.34E-02	1.00E+00
GO:0065008	regulation of biological quality	13	5.22E-04	4.34E-02	1.00E+00
GO:0051239	regulation of multicellular organismal process	11	5.23E-04	4.34E-02	1.00E+00
GO:0060337	type I interferon signaling pathway	3	5.30E-04	4.34E-02	1.00E+00
GO:0050728	negative regulation of inflammatory response	3	5.30E-04	4.34E-02	1.00E+00
GO:0071357	cellular response to type I interferon	3	5.30E-04	4.34E-02	1.00E+00
GO:0042176	regulation of protein catabolic process	4	5.47E-04	4.44E-02	1.00E+00
GO:0034340	response to type I interferon	3	5.50E-04	4.44E-02	1.00E+00
GO:0071496	cellular response to external stimulus	4	5.69E-04	4.57E-02	1.00E+00
GO:0002820	negative regulation of adaptive immune response	2	5.85E-04	4.67E-02	1.00E+00
GO:0034097	response to cytokine	6	5.93E-04	4.70E-02	1.00E+00
GO:0032101	regulation of response to external stimulus	6	6.17E-04	4.86E-02	1.00E+00
GO:0051241	negative regulation of multicellular organismal process	5	6.29E-04	4.93E-02	1.00E+00

missSNP set **MF terms**

GO_id		No of genes	p-value	p-value_fdr	p-value_bonferroni
GO:0005126	cytokine receptor binding	5	6.69E-07	2.57E-03	2.57E-03
GO:0005138	interleukin-6 receptor binding	2	9.09E-06	1.74E-02	3.49E-02
GO:0050700	CARD domain binding	2	2.54E-05	3.25E-02	9.75E-02

ncSNP set none

allSNP set

GO:0005126	cytokine receptor binding	7	5.80E-08	2.23E-04	2.23E-04
GO:0004918	interleukin-8 receptor activity	2	3.90E-06	7.49E-03	1.50E-02
GO:0005138	interleukin-6 receptor binding	2	3.89E-05	4.20E-02	1.49E-01
GO:0070851	growth factor receptor binding	4	4.38E-05	4.20E-02	1.68E-01
GO:0050700	CARD domain binding	2	1.08E-04	8.33E-02	4.16E-01

missSNP set **CC terms**

GO_id		NumberOfGenes	p-value	p-value_fdr	p-value_bonferroni
GO:0009986	cell surface	6	1.26E-05	1.86E-02	1.86E-02

ncSNP set none
allSNP set none

Table 22. GO terms enrichemnt for missSNP geneset by ConsensusPathDB

Enrichement of GO BP terms for missSNP gene set

uploaded list: 23
mapped entities: 22
enriched gene ontology-based sets: 26
21 genes (95.5%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0006955	immune response	BP 2	1457	13 (0.9%)	4.43E-10	2.26E-08
GO:0002252	immune effector process	BP 2	631	9 (1.5%)	1.19E-08	3.04E-07
GO:0051707	response to other organism	BP 2	744	8 (1.1%)	8.18E-07	1.39E-05
GO:0009607	response to biotic stimulus	BP 2	776	8 (1.1%)	1.13E-06	1.45E-05
GO:0045321	leukocyte activation	BP 2	673	7 (1.1%)	5.82E-06	5.94E-05
GO:0001913	T cell mediated cytotoxicity	BP 2	37	3 (8.6%)	8.68E-06	7.38E-05
GO:0002440	production of molecular mediator of immune response	BP 2	130	4 (3.2%)	1.24E-05	9.02E-05
GO:0002520	immune system development	BP 2	786	7 (0.9%)	1.65E-05	0.000105
GO:0006950	response to stress	BP 2	3557	13 (0.4%)	2.03E-05	0.000115
GO:0001909	leukocyte mediated cytotoxicity	BP 2	78	3 (3.9%)	9.04E-05	0.000454
GO:0008283	cell proliferation	BP 2	1861	9 (0.5%)	9.78E-05	0.000454
GO:0002253	activation of immune response	BP 2	438	5 (1.2%)	0.000108	0.000458
GO:0051716	cellular response to stimulus	BP 2	6497	16 (0.3%)	0.000134	0.000527
GO:0009605	response to external stimulus	BP 2	2113	9 (0.4%)	0.000258	0.000938
GO:0044700	single organism signaling	BP 2	5897	14 (0.2%)	0.00102	0.00339
GO:0098602	single organism cell adhesion	BP 2	721	5 (0.7%)	0.00106	0.00339
GO:0019882	antigen processing and presentation	BP 2	233	3 (1.3%)	0.00235	0.00706
GO:0044767	single-organism developmental process	BP 2	5402	12 (0.2%)	0.00696	0.0197
GO:0044763	single-organism cellular process	BP 2	11949	19 (0.2%)	0.00814	0.0219

Enrichement of GO MF terms for missSNP set

uploaded list: 23
mapped entities: 22
enriched gene ontology-based sets: 8
21 genes (95.5%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0005126	cytokine receptor binding	MF 4	257	5 (2.0%)	8.01E-06	0.000112
GO:0050700	CARD domain binding	MF 4	8	2 (25.0%)	3.60E-05	0.000252
GO:0004896	cytokine receptor activity	MF 4	90	3 (3.8%)	0.000105	0.000492
GO:0070851	growth factor receptor binding	MF 4	122	3 (2.5%)	0.000333	0.00116
GO:0005515	protein binding	MF 2	8713	17 (0.2%)	0.00169	0.0203
GO:0005102	receptor binding	MF 3	1369	6 (0.5%)	0.00331	0.0238
GO:0019955	cytokine binding	MF 3	80	2 (2.7%)	0.0034	0.0238
GO:0003823	antigen binding	MF 2	108	2 (2.0%)	0.0062	0.0372

Enrichement of GO CC terms for missSNP set

uploaded list: 23
mapped entities: 22
enriched gene ontology-based sets: 5
21 genes (95.5%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0009986	cell surface	CC 2	690	5 (0.8%)	0.0008	0.0176
GO:0044459	plasma membrane part	CC 2	2279	8 (0.4%)	0.00206	0.0226
GO:0098552	side of membrane	CC 2	307	3 (1.0%)	0.00433	0.0317
GO:0005886	plasma membrane	CC 2	4776	11 (0.2%)	0.00765	0.0396
GO:0071944	cell periphery	CC 2	4870	11 (0.2%)	0.00899	0.0396

Table 23. Enrichment of BP, MF and CC GO terms for ncSNP geneset

uploaded list:	23
mapped entities:	23
enriched gene ontology-based sets:	36

20 genes (87.0%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0006955	immune response	BP 2	1457	10 (0.7%)	7.49E-07	3.82E-05
GO:0045321	leukocyte activation	BP 2	673	7 (1.1%)	4.00E-06	0.000102
GO:0002520	immune system development	BP 2	786	7 (0.9%)	1.14E-05	0.000195
GO:0050900	leukocyte migration	BP 2	321	5 (1.6%)	1.76E-05	0.000225
GO:0009605	response to external stimulus	BP 2	2113	10 (0.5%)	2.24E-05	0.000229
GO:0006950	response to stress	BP 2	3557	12 (0.3%)	6.64E-05	0.000564
GO:0042221	response to chemical	BP 2	3872	12 (0.3%)	0.00015	0.000873
GO:0048870	cell motility	BP 2	1175	7 (0.6%)	0.000154	0.000873
GO:0051674	localization of cell	BP 2	1175	7 (0.6%)	0.000154	0.000873
GO:0008283	cell proliferation	BP 2	1861	8 (0.4%)	0.000425	0.00195
GO:0044765	single-organism transport	BP 2	3605	11 (0.3%)	0.000456	0.00195
GO:0002252	immune effector process	BP 2	631	5 (0.8%)	0.000458	0.00195
GO:0042330	taxis	BP 2	692	5 (0.7%)	0.000682	0.00261
GO:1902578	single-organism localization	BP 2	3791	11 (0.3%)	0.000716	0.00261
GO:0098602	single organism cell adhesion	BP 2	721	5 (0.7%)	0.000837	0.00284
GO:0065008	regulation of biological quality	BP 2	3239	10 (0.3%)	0.000892	0.00284
GO:0051707	response to other organism	BP 2	744	5 (0.7%)	0.000956	0.00287
GO:0009607	response to biotic stimulus	BP 2	776	5 (0.7%)	0.00116	0.00329
GO:0051234	establishment of localization	BP 2	4397	11 (0.3%)	0.00263	0.00705
GO:0001776	leukocyte homeostasis	BP 2	75	2 (2.8%)	0.00285	0.00726
GO:0050789	regulation of biological process	BP 2	10286	17 (0.2%)	0.00571	0.0131
GO:0051716	cellular response to stimulus	BP 2	6497	13 (0.2%)	0.00582	0.0131
GO:0044707	single-multicellular organism process	BP 2	6462	13 (0.2%)	0.0059	0.0131
GO:0044459	plasma membrane part	CC 2	2279	7 (0.3%)	0.00672	0.128
GO:0044700	single organism signaling	BP 2	5897	12 (0.2%)	0.0082	0.0174
GO:0033036	macromolecule localization	BP 2	2313	7 (0.3%)	0.00877	0.0179
GO:0002253	activation of immune response	BP 2	438	3 (0.7%)	0.0111	0.0218
GO:0016265	death	BP 2	1899	6 (0.3%)	0.0126	0.0238
GO:0007155	cell adhesion	BP 2	1375	5 (0.4%)	0.0138	0.0251
GO:0009719	response to endogenous stimulus	BP 2	1409	5 (0.4%)	0.0144	0.0254
GO:0004871	signal transducer activity	MF 2	1687	5 (0.3%)	0.0263	0.162
GO:0003700	sequence-specific DNA binding transcription factor activity	MF 2	1090	4 (0.4%)	0.0279	0.162
GO:0005515	protein binding	MF 2	8713	14 (0.2%)	0.0346	0.162
GO:0044767	single-organism developmental process	BP 2	5402	10 (0.2%)	0.0405	0.0689
GO:0009653	anatomical structure morphogenesis	BP 2	2484	6 (0.2%)	0.0435	0.0715
GO:0005886	plasma membrane	CC 2	4776	9 (0.2%)	0.0466	0.332

Table 24. Enrichment of BP, MF and CC GO terms for allSNP geneset

Enriched gene ontology-based sets for allSNP set for BP GO

uploaded list: 56
mapped entities: 54
enriched gene ontology-based sets: 44
41 genes (91.1%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0006955	immune response	BP 2	1457	23 (1.6%)	1.56E-15	9.33E-14
GO:0002252	immune effector process	BP 2	631	14 (2.3%)	3.72E-11	1.12E-09
GO:0045321	leukocyte activation	BP 2	673	14 (2.1%)	8.05E-11	1.61E-09
GO:0002520	immune system development	BP 2	786	14 (1.8%)	6.55E-10	9.82E-09
GO:0051707	response to other organism	BP 2	744	13 (1.8%)	3.85E-09	4.59E-08
GO:0006950	response to stress	BP 2	3557	25 (0.7%)	4.59E-09	4.59E-08
GO:0009607	response to biotic stimulus	BP 2	776	13 (1.7%)	6.48E-09	5.55E-08
GO:0009605	response to external stimulus	BP 2	2113	19 (0.9%)	2.21E-08	1.66E-07
GO:0008283	cell proliferation	BP 2	1861	17 (0.9%)	1.45E-07	9.64E-07
GO:0098602	single organism cell adhesion	BP 2	721	10 (1.4%)	2.80E-06	1.68E-05
GO:0051716	cellular response to stimulus	BP 2	6497	29 (0.5%)	3.10E-06	1.69E-05
GO:0002253	activation of immune response	BP 2	438	8 (1.9%)	4.44E-06	2.22E-05
GO:0050900	leukocyte migration	BP 2	321	7 (2.3%)	5.39E-06	2.49E-05
GO:0002440	production of molecular mediator of immune	BP 2	130	5 (4.0%)	9.40E-06	4.03E-05
GO:0044700	single organism signaling	BP 2	5897	26 (0.5%)	2.77E-05	0.000111
GO:0048870	cell motility	BP 2	1175	11 (1.0%)	3.54E-05	0.000125
GO:0051674	localization of cell	BP 2	1175	11 (1.0%)	3.54E-05	0.000125
GO:0044459	plasma membrane part	CC 2	2279	15 (0.7%)	4.47E-05	0.0013
GO:0042221	response to chemical	BP 2	3872	20 (0.5%)	6.07E-05	0.000202
GO:0001913	T cell mediated cytotoxicity	BP 2	37	3 (8.6%)	6.78E-05	0.000214
GO:0097278	complement-dependent cytotoxicity	BP 2	6	2 (33.3%)	7.51E-05	0.000225
GO:0044707	single-multicellular organism process	BP 2	6462	26 (0.4%)	0.000186	0.000531
GO:0005515	protein binding	MF 2	8713	31 (0.4%)	0.00021	0.00357
GO:0050789	regulation of biological process	BP 2	10286	34 (0.3%)	0.000213	0.000581
GO:0065008	regulation of biological quality	BP 2	3239	17 (0.5%)	0.000277	0.000723
GO:0001776	leukocyte homeostasis	BP 2	75	3 (4.2%)	0.000583	0.00145
GO:0016265	death	BP 2	1899	12 (0.6%)	0.000604	0.00145
GO:0009986	cell surface	CC 2	690	7 (1.1%)	0.00064	0.00927
GO:0007155	cell adhesion	BP 2	1375	10 (0.7%)	0.000678	0.00152
GO:0001909	leukocyte mediated cytotoxicity	BP 2	78	3 (3.9%)	0.000682	0.00152
GO:0042330	taxis	BP 2	692	7 (1.0%)	0.000712	0.00153
GO:0044767	single-organism developmental process	BP 2	5402	22 (0.4%)	0.000897	0.00186
GO:0044110	growth involved in symbiotic interaction	BP 2	21	2 (9.5%)	0.00103	0.00206
GO:0033036	macromolecule localization	BP 2	2313	13 (0.6%)	0.00113	0.0022
GO:0005886	plasma membrane	CC 2	4776	20 (0.4%)	0.00114	0.0108
GO:0071944	cell periphery	CC 2	4870	20 (0.4%)	0.00149	0.0108
GO:0009719	response to endogenous stimulus	BP 2	1409	9 (0.7%)	0.00297	0.00542
GO:0044763	single-organism cellular process	BP 2	11949	35 (0.3%)	0.00299	0.00542
GO:0044765	single-organism transport	BP 2	3605	16 (0.5%)	0.00307	0.00542

GO:0051234	establishment of localization	BP 2	4397	18 (0.4%)	0.0037	0.00634
GO:0098552	side of membrane	CC 2	307	4 (1.4%)	0.00405	0.0235
GO:0048856	anatomical structure development	BP 2	4828	19 (0.4%)	0.00419	0.00698
GO:1902578	single-organism localization	BP 2	3791	16 (0.4%)	0.00517	0.00838
GO:0004871	signal transducer activity	MF 2	1687	9 (0.6%)	0.00805	0.0685

Enriched gene ontology-based sets for allSNP set for MF GO

uploaded list: 56
mapped entities: 54
enriched gene ontology-based sets: 12
41 genes (91.1%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0004896	cytokine receptor activity	MF 4	90	6 (7.5%)	2.49E-08	7.98E-07
GO:0005126	cytokine receptor binding	MF 4	257	7 (2.8%)	1.29E-06	2.06E-05
GO:0019955	cytokine binding	MF 3	80	4 (5.3%)	2.48E-05	0.000546
GO:0050700	CARD domain binding	MF 4	8	2 (25.0%)	0.00014	0.00117
GO:0070851	growth factor receptor binding	MF 4	122	4 (3.4%)	0.000146	0.00117
GO:0005515	protein binding	MF 2	8713	31 (0.4%)	0.00021	0.00357
GO:0019956	chemokine binding	MF 4	13	2 (16.7%)	0.000328	0.0021
GO:0008528	G-protein coupled peptide recept	MF 4	121	3 (2.9%)	0.00174	0.00839
GO:0001653	peptide receptor activity	MF 4	123	3 (2.8%)	0.00183	0.00839
GO:0005102	receptor binding	MF 3	1369	9 (0.7%)	0.00257	0.0282
GO:0004871	signal transducer activity	MF 2	1687	9 (0.6%)	0.00805	0.0685
GO:0019904	protein domain specific binding	MF 3	586	5 (0.9%)	0.00893	0.0655

Enriched gene ontology-based sets for allSNP set for CC GO

uploaded list: 56
mapped entities: 54
enriched gene ontology-based sets: 7
41 genes (91.1%) from the input list are present in at least one GO category.

gene ontology term		category, level	set size	candidates contained	p-value	q-value
GO:0044459	plasma membrane part	CC 2	2279	15 (0.7%)	4.47E-05	0.0013
GO:0009986	cell surface	CC 2	690	7 (1.1%)	0.00064	0.00927
GO:0005886	plasma membrane	CC 2	4776	20 (0.4%)	0.00114	0.0108
GO:0071944	cell periphery	CC 2	4870	20 (0.4%)	0.00149	0.0108
GO:0005887	integral component of plasma me	CC 4	1342	9 (0.7%)	0.00172	0.031
GO:0031226	intrinsic component of plasma me	CC 3	1396	9 (0.7%)	0.00229	0.0481
GO:0098552	side of membrane	CC 2	307	4 (1.4%)	0.00405	0.0235

Table 25.

DiseaseConnectDB KEGG pathway dataset for AID: pleiotropy at the pathway-disease level

Pathway map/module		Pathway name	Pathway ID	AID Diseases
Environmental Information Processing	Signal molecules	Cytokine-cytokine receptor interaction	hsa04060	RA CD AS
	Signal transduction	NF-kappa B signaling pathway	hsa04064	RA CD PS
	Signal transduction	Jak-STAT signaling pathway	hsa04630	RA CD PS
Organismal Systems	Immune system	Antigen processing and presentation	hsa04612	RA CD
	Immune system	NOD-like receptor signaling pathway	hsa04621	CD PS
	Immune system	T cell receptor signaling pathway	hsa04660	RA CD
	Immune system	Intestinal immune network for IgA production	hsa04672	RA CD
Human Diseases				
	Endocrine and metabolic diseases	Type I diabetes mellitus	hsa04940	CD PS
	Immune diseases	Asthma	hsa05310	RA CD
		Autoimmune thyroid disease	hsa05320	RA CD
		Systemic lupus erythematosus	hsa05322	RA CD
		Rheumatoid arthritis	hsa05323	RA CD
		Graft-versus-host disease	hsa05332	RA CD
		Allograft rejection	hsa05330	RA CD PsA
	Infectious diseases			
	Infectious diseases: Parasitic			
		Malaria	hsa05144	RA CD
		Toxoplasmosis	hsa05145	RA CD PS
		Amoebiasis	hsa05146	RA CD AS
	Infectious diseases: Bacterial			
		Staphylococcus aureus infection	hsa05150	RA CD
		Tuberculosis	hsa05152	RA CD
	Infectious diseases: Viral			
		Measles	hsa05162	RA PS
		Influenza A	hsa05164	RA CD PS
		HTLV-I infection	hsa05166	RA CD
		Herpes simplex infection	hsa05168	RA CD PsA PS

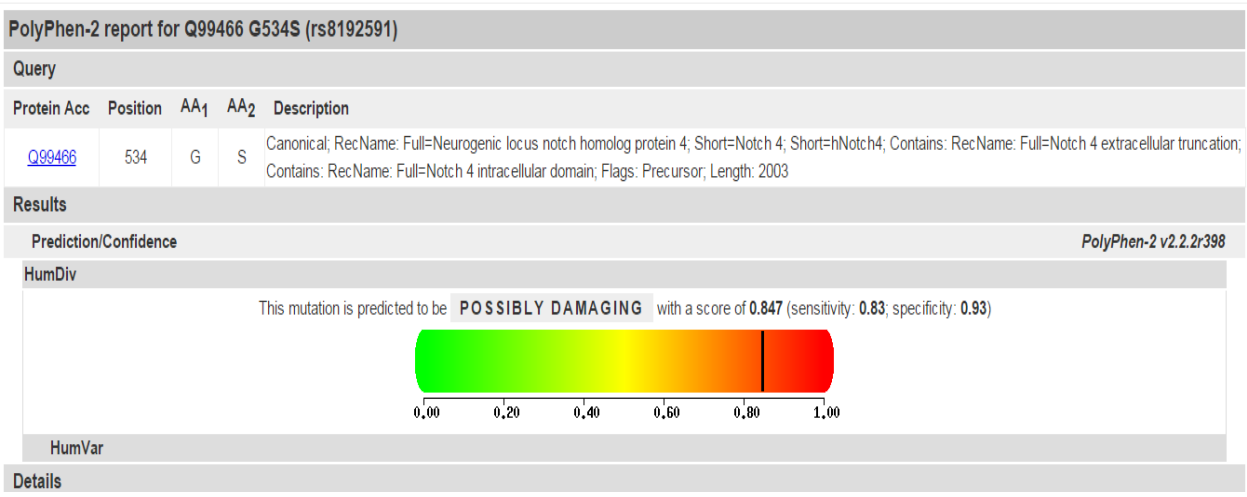
Source is Disease ConnectDB; by each AID:
https://docs.google.com/spreadsheets/d/1kMUw7xJHXjYXrXglWsCWDQ6qcrTlY8L1QSHY3BJlJ9k/edit#gid=2121775356

Table 26. STRING disease enrichment data for AID GWAS SNP harboring geneset

GO_id	Term	No of Genes	p-value	p-value_fdr	p-value_ bonferroni
DOID:8778	**Crohn's disease**	3	3.23E-09	6.20E-06	1.49E-05
DOID:0050589	**Inflammatory bowel disease**	3	3.23E-09	6.20E-06	1.49E-05
DOID:3342	Bone inflammation disease	4	4.04E-09	6.20E-06	1.86E-05
DOID:9008	**Psoriatic arthritis**	3	4.52E-08	5.19E-05	2.08E-04
DOID:225	Syndrome	3	9.66E-08	8.89E-05	4.45E-04
DOID:5295	Intestinal disease	3	6.54E-07	4.87E-04	3.01E-03
DOID:4	Disease	11	7.41E-07	4.87E-04	3.41E-03
DOID:0080001	Bone disease	4	1.31E-06	7.55E-04	6.04E-03
DOID:7147	**Ankylosing spondylitis**	2	5.46E-06	2.79E-03	2.51E-02
DOID:7	Disease of anatomical entity	8	7.07E-06	3.25E-03	3.25E-02
DOID:77	Gastrointestinal system disease	3	8.42E-06	3.52E-03	3.87E-02
DOID:17	Musculoskeletal system disease	4	5.64E-05	2.16E-02	2.60E-01
DOID:676	**Juvenile rheumatoid arthritis**	2	9.49E-05	3.12E-02	4.37E-01
DOID:7148	**Rheumatoid arthritis**	2	9.49E-05	3.12E-02	4.37E-01
DOID:848	Arthritis	2	1.23E-04	3.77E-02	5.65E-01

Diseases IDs and names are given, in bold are emphasized AIDs; number of genes in a gene set for each disease; p values, after fdr and Bonferroni corrections.

Figure 1. An example of functional evaluation of missense SNP impact by PolyPhen-2



Prediction of conformational alteration caused by an amino acid change: the resulting report for rs8192591 is presented with a score on the scale from zero to one and color coded green to red for a level of damage. Specificity and sensitivity of evaluation for each test is also provided by PolyPhen-2.

Figure 2a: TNF network

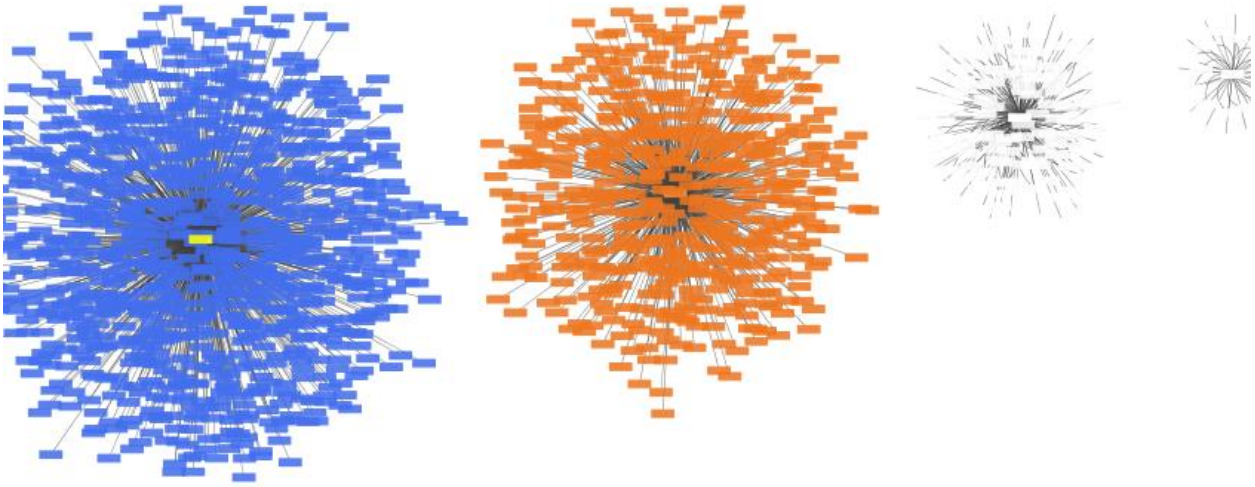
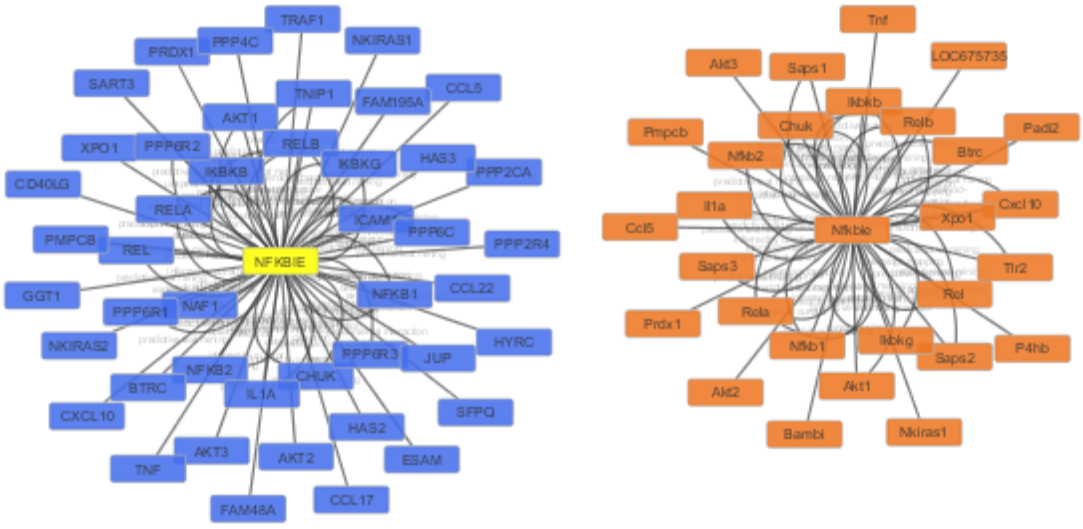


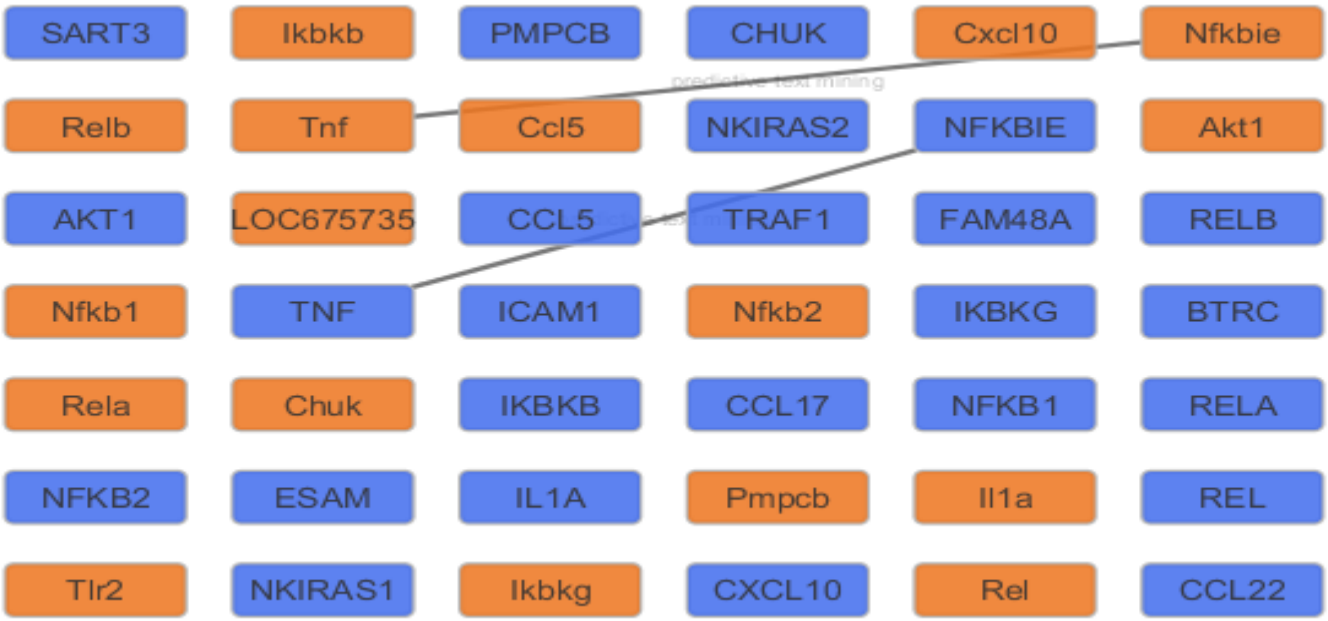
Figure 2b: NFKBIE network



Networks for SNP genes and TNF are presented: color coded for a species (human is blue, red is murine network, white is for rat); nodes represent proteins; edges represent any type of interactions between two proteins.

Figure 3a. Image of intersection between ERAP1 and TNF networks (nodes represent proteins that are common to both networks)

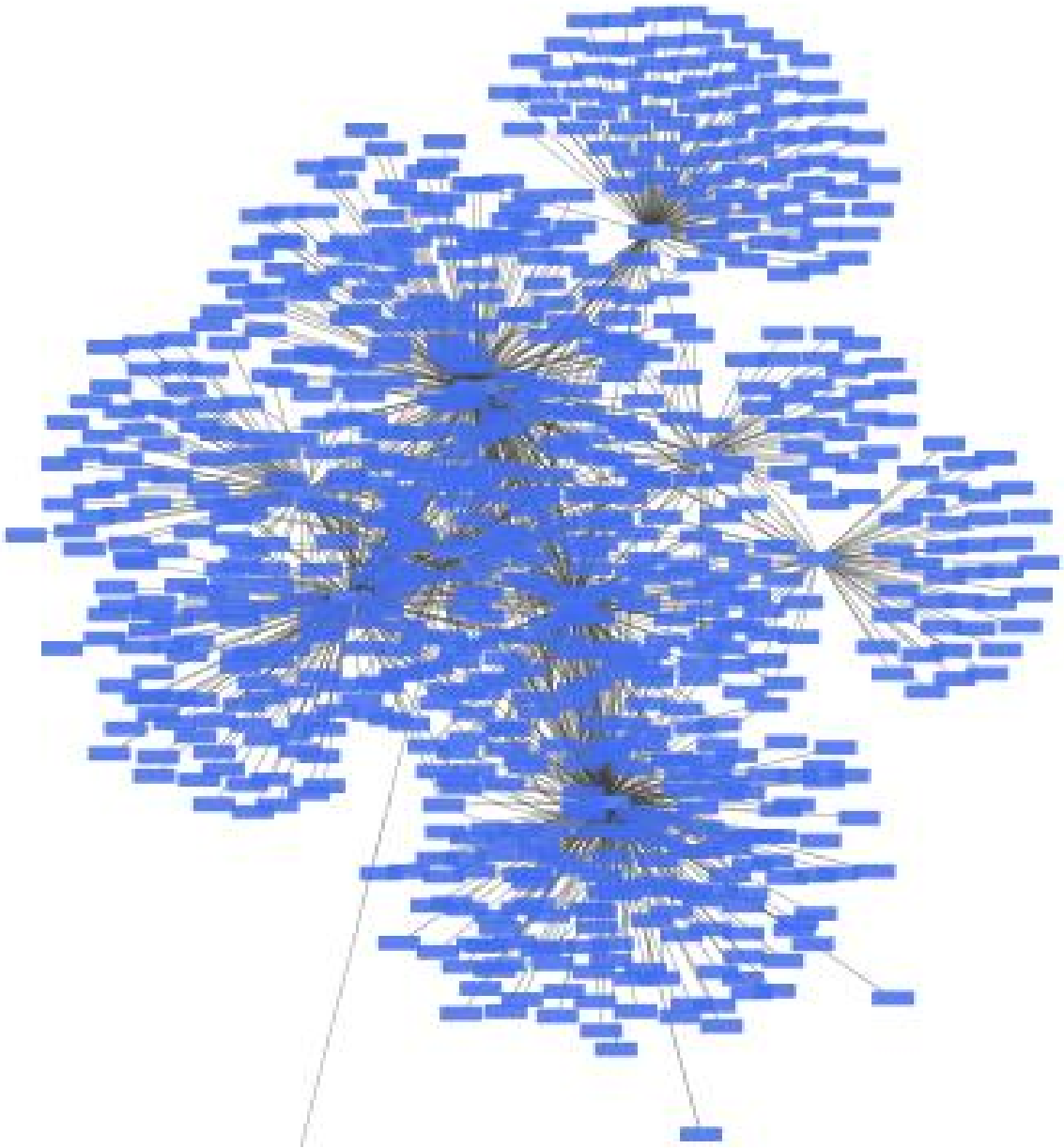
Figure 3b. Image of intersection between missSNP NFKBIE and TNF networks
(nodes represent proteins that are common to both networks)



Nodes labeled with gene symbols; color-coded for each species: blue for human, and red and white for mouse and rat respectively.

Figure 4.

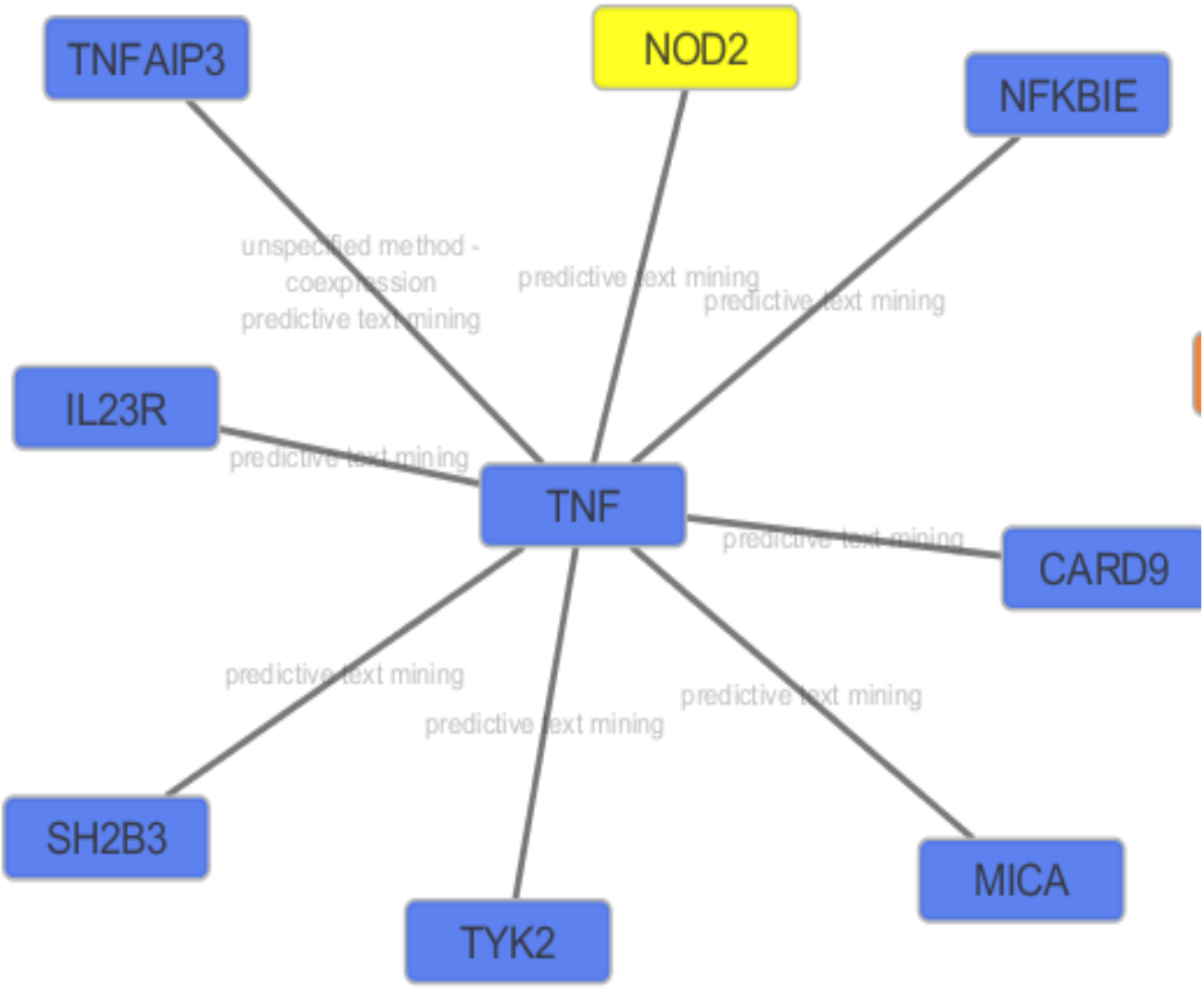
Union of GWAS AID SNP harboring protein networks



Nodes labeled with official gene symbols; edges represent links between pairs of nodes; clouds are color-coded blue for human.

Figure 5.

Network union between intersection datasets of all missSNPs and TNF networks



Resulting network consists of the nodes labeled with official human gene symbols and edges represent links between nodes (proteins) that were detected by different methods such as predictive textmining and coexpression.

Figure 6.

GWAS AID missSNP harboring genes show high expression in several immune-related human tissues and cells

colored squares indicate intensity of expression for each gene and tissues/cells

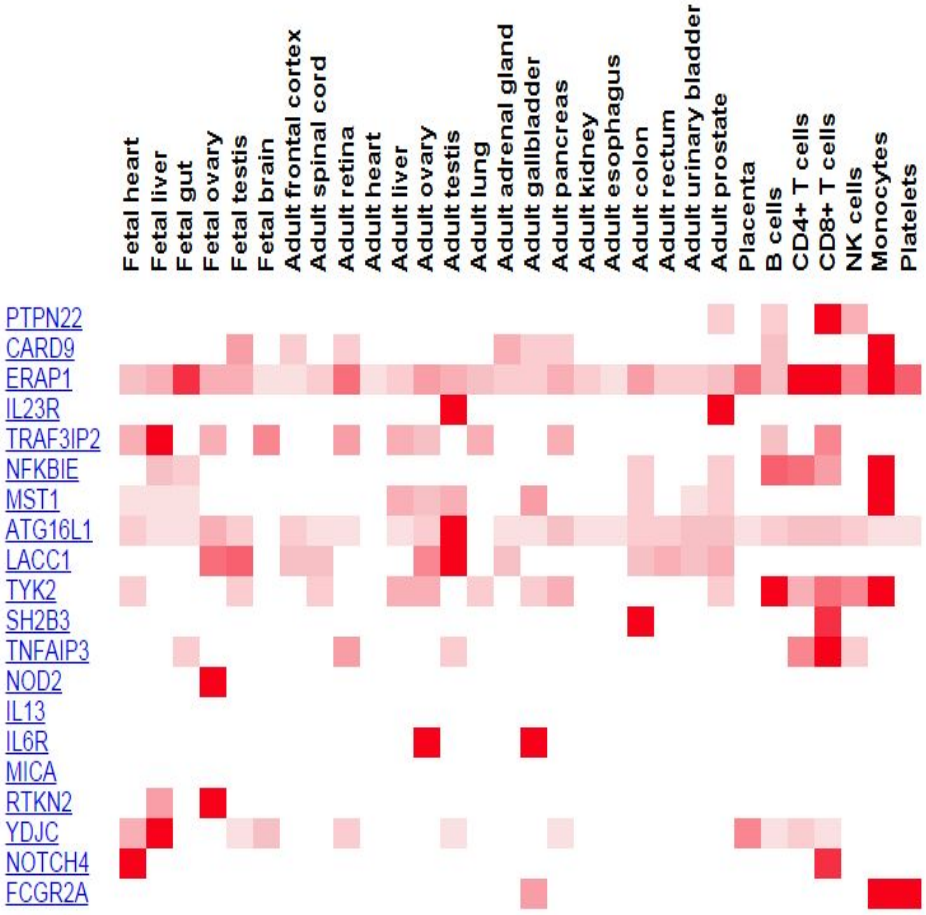
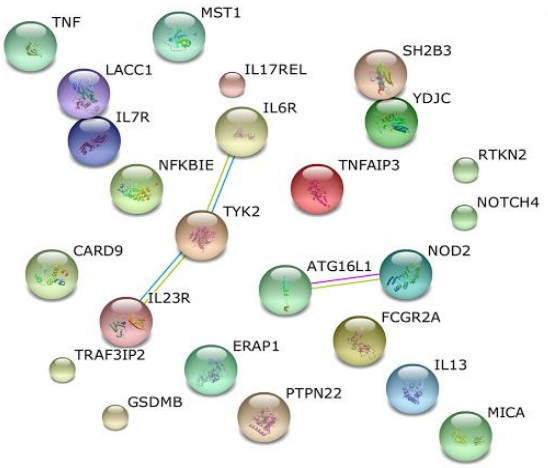
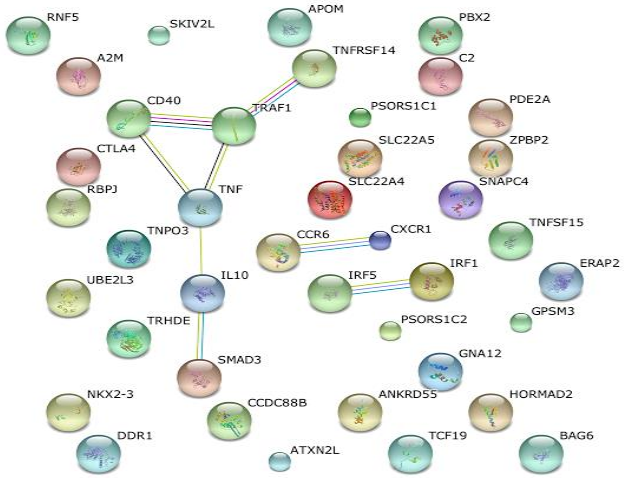


Figure 7. Image of networks constructed by STRING for missSNP, ncSNP and allSNP datasets

7a) missSNP



7b) ncSNP



7c) allSNP

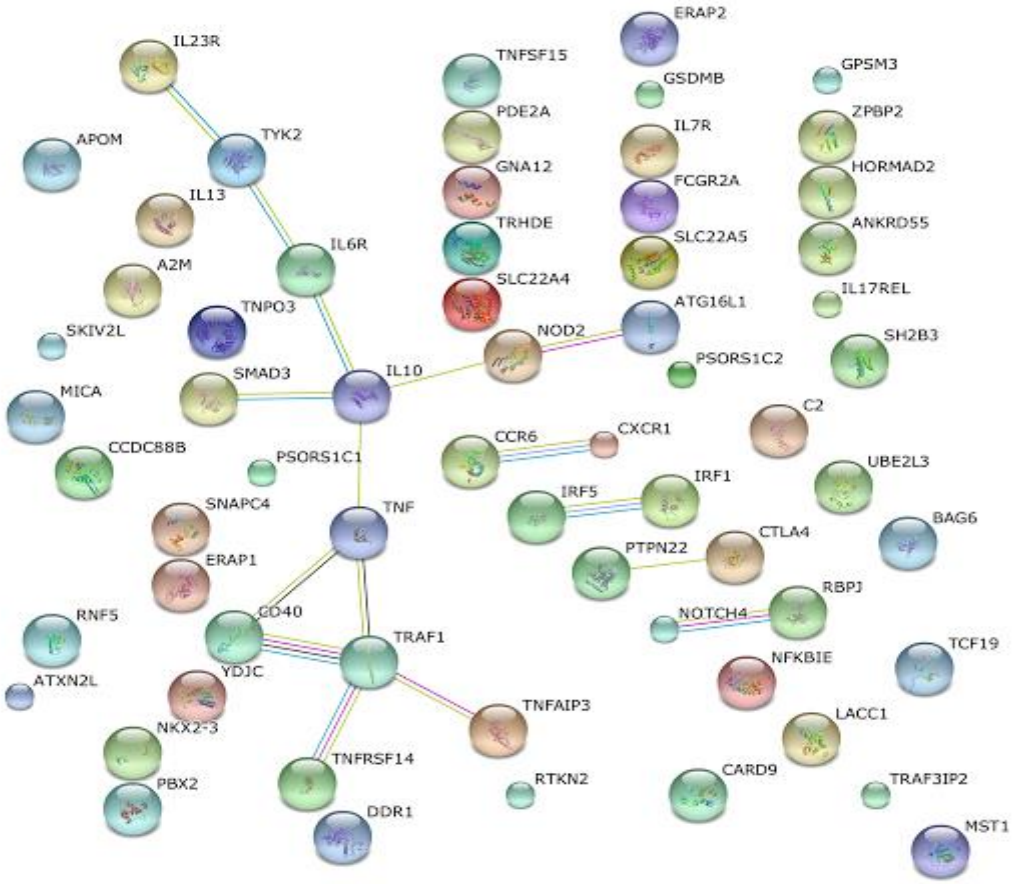
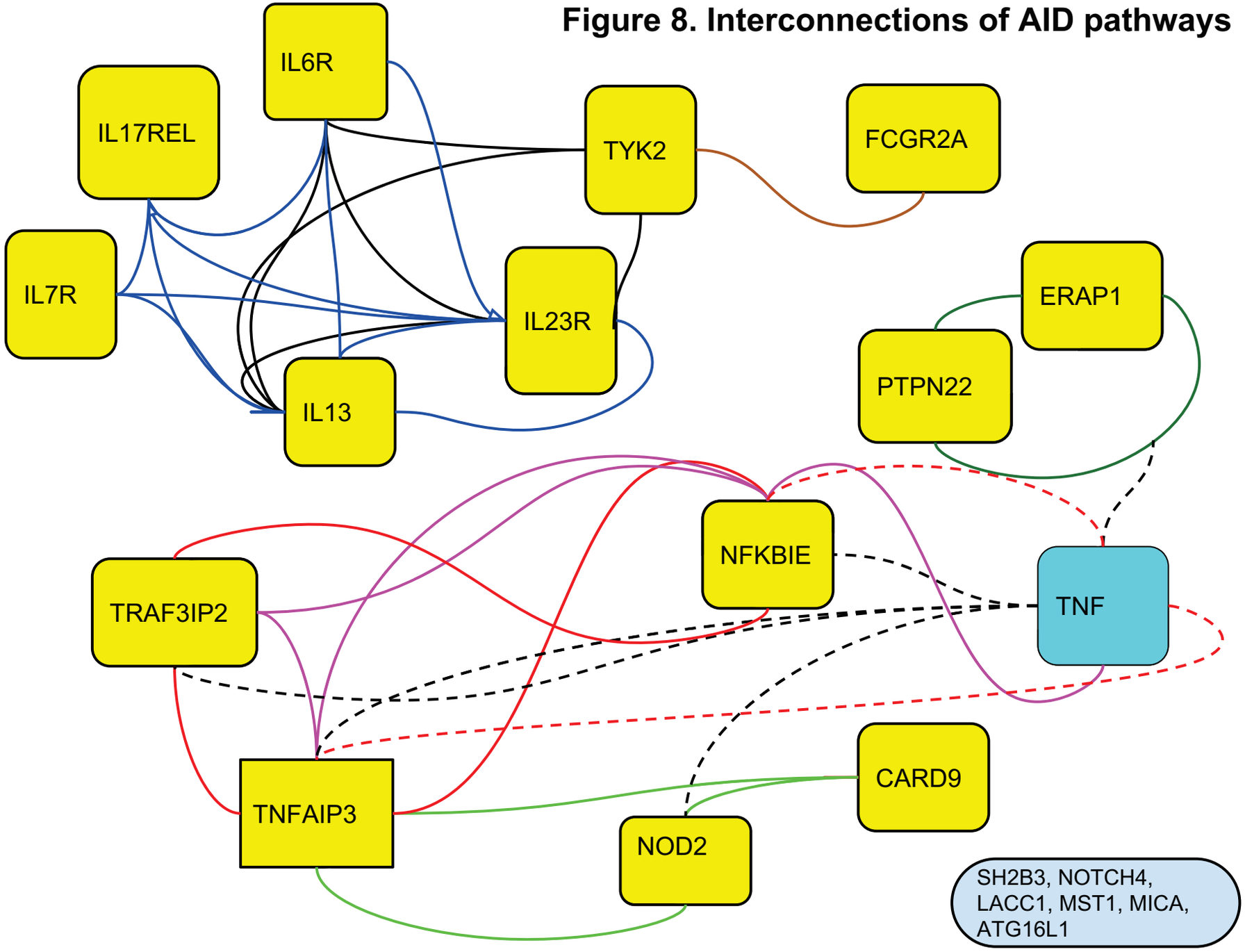


Figure 8. Interconnections of AID pathways



Legend: connecting lines between genes denote KEGG pathways:

red: NFkB signaling

black: Jak_STAT signaling

blue: Cytokine-Cytokine R

green: NOD-like signaling

brown: Osteoclast differentiation

pink: TNF signaling

dark green: Antigen Proc&Presen

black dotted: TNF signaling

red dotted: T cell R signaling

Color of gene's cell:

<u>yellow background:</u>
gene is connected to the selected KEGG pathways

<u>blue background:</u>
gene is not connected to the selected KEGG pathways

Supplemental Table 1. AID GWAS associations retrieved from NHGRI Catalog

at the end if November 201

Arthritis, Psoriatic PsA

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	PubMed
1	rs13191343	nearGene-5	HLA-C	6 : 31,241,109	2.000 x 10-72	4,5,6	NHGRI	20953186
2	rs33980500	missense	TRAF3IP2	6 : 111,913,262	1.000 x 10-20	"	NHGRI	20953186
3	rs12188300	intergenic	IL12B, ADRA1B	5 : 158,829,527	7.000 x 10-17	"	NHGRI	20953186
4	rs13017599	intergenic	REL, RPS12P3	2 : 61,164,331	1.000 x 10-8	"	NHGRI	22170493
5	rs702873	intergenic	RPL21P33, REL	2 : 61,081,542	2.000 x 10-7	"	NHGRI	22170493

Arthritis, Rheumatoid RA

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	PubMed
1	rs4750316	UTR-3	DKFZp667F0711	10 : 6,393,260	2.000 x 10-7	*	NHGRI	20453842
2	rs4750316	UTR-3	DKFZp667F0711	10 : 6,393,260	4.000 x 10-7	*	NHGRI	18794853
3	rs1329568	UTR-3	LOC100130458	9 : 37,037,976	8.000 x 10-7	*	NHGRI	22491018
4	rs805297	nearGene-5	APOM	6 : 31,622,606	3.000 x 10-10	**1f**	NHGRI	21844665
5	rs2812378	nearGene-5	CCL21	9 : 34,710,260	3.000 x 10-8	*	NHGRI	18794853
6	rs2841277	nearGene-5	PLD4	14 : 105,391,005	2.000 x 10-14	*	NHGRI	22446963
7	rs6496667	nearGene-5	ZNF774	15 : 90,893,668	1.000 x 10-7	*	NHGRI	22446963
8	rs3087243	nearGene-3	CTLA4	2 : 204,738,919	1.000 x 10-8	**3a**	NHGRI	20453842
9	rs10488631	nearGene-3	TNPO3	7 : 128,594,183	4.000 x 10-11	**3a**	NHGRI	20453842
10	rs2233434	missense	NFKBIE	6 : 44,232,920	6.000 x 10-19	*	NHGRI	22446963
11	rs2476601	missense	PTPN22	1 : 114,377,568	9.000 x 10-74	**2b**	NHGRI	20453842
12	rs2476601	missense	PTPN22	1 : 114,377,568	2.000 x 10-21	**2b**	NHGRI	19503088
13	rs2476601	missense	PTPN22	1 : 114,377,568	2.000 x 10-11	**2b**	NHGRI	17804836
14	rs3184504	missense	SH2B3	12 : 111,884,608	6.000 x 10-7	*	NHGRI	20453842
15	rs2230926	missense	TNFAIP3	6 : 138,196,066	2.000 x 10-7	*	NHGRI	20453841
16	rs2075876	intron	AIRE	21 : 45,709,153	4.000 x 10-9	*	NHGRI	21505073
18	rs6859219	intron	ANKRD55	5 : 55,438,580	1.000 x 10-11	**2b**	NHGRI	20453842
19	rs2867461	intron	ANXA3	4 : 79,513,215	1.000 x 10-12	*	NHGRI	22446963
21	rs10821944	intron	ARID5B	10 : 63,785,089	6.000 x 10-18	*	NHGRI	22446963
23	rs26232	intron	C5orf30	5 : 102,596,720	4.000 x 10-8	**2b**	NHGRI	20453842
24	rs6910071	intron	C6orf10	6 : 32,282,854	1.000 x 10-299	*	NHGRI	20453842
25	rs2395148	intron	C6orf10	6 : 32,321,554	2.000 x 10-10	*	NHGRI	18576341
26	rs3093024	intron	CCR6	6 : 167,532,793	8.000 x 10-19	*	NHGRI	20453841
27	rs3093023	intron	CCR6	6 : 167,534,290	2.000 x 10-11	**1f**	NHGRI	20453842
28	rs840016	intron	CD247	1 : 167,408,670	2.000 x 10-7	*	NHGRI	20453842
29	rs4810485	intron	CD40	20 : 44,747,947	3.000 x 10-9	**1f**	NHGRI	20453842
30	rs4810485	intron	CD40	20 : 44,747,947	8.000 x 10-9	**1f**	NHGRI	18794853
31	rs42041	intron	CDK6	7 : 92,246,744	4.000 x 10-7	*	NHGRI	18794853
32	rs4942242	intron	ENOX1	13 : 44,217,064	2.000 x 10-7	*	NHGRI	22491018
33	rs1914816	intron	ETFA	15 : 76,546,933	7.000 x 10-7	*	NHGRI	22491018
34	rs13315591	intron	FAM107A	3 : 58,556,841	5.000 x 10-8	*	NHGRI	20453842
35	rs7940423	intron	GALNTL4	11 : 11,504,228	1.000 x 10-7	*	NHGRI	22491018
36	rs3783637	intron	GCH1	14 : 55,348,118	2.000 x 10-7	**2b**	NHGRI	22446963
37	rs706778	intron	IL2RA	10 : 6,098,949	1.000 x 10-11	*	NHGRI	20453842
38	rs13119723	intron	KIAA1109	4 : 123,218,313	7.000 x 10-7	*	NHGRI	20453842
39	rs1678542	intron	KIF5A	12 : 57,968,715	9.000 x 10-8	*	NHGRI	18794853
40	rs13393173	intron	LASS6	2 : 169,389,091	4.000 x 10-7	*	NHGRI	18615156
41	rs17118552	intron	MDGA2	14 : 47,874,557	2.000 x 10-7	*	NHGRI	22491018
42	rs3890745	intron	MMEL1	1 : 2,553,624	1.000 x 10-7	*	NHGRI	18794853
43	rs3890745	intron	MMEL1	1 : 2,553,624	4.000 x 10-7	*	NHGRI	20453842
44	rs7046653	intron	MOBKL2B	9 : 27,490,967	5.000 x 10-7	*	NHGRI	18615156
45	rs6500395	intron	N4BP1	16 : 48,621,402	6.000 x 10-7	*	NHGRI	22491018

46	rs3781913	intron	PDE2A	11	: 72,373,496	6.000 x 10-10	**1f**	NHGRI	22446963
47	rs2075876	intron	PFKL	21	: 45,709,153	4.000 x 10-9	*	NHGRI	21505073
48	rs6026990	intron	PHACTR3	20	: 58,177,615	6.000 x 10-7	*	NHGRI	22491018
49	rs854555	intron	PON1	7	: 94,930,391	2.000 x 10-7	*	NHGRI	18615156
50	rs7404928	intron	PRKCB	16	: 23,888,840	4.000 x 10-7	*	NHGRI	22446963
51	rs1957895	intron	PRKCH	14	: 61,908,332	4.000 x 10-7	*	NHGRI	22446963
52	rs12901682	intron	PSMA4	15	: 78,833,223	4.000 x 10-8	*	NHGRI	22491018
53	rs2847297	intron	PTPN2	18	: 12,797,694	2.000 x 10-8	*	NHGRI	22446963
54	rs13137105	intron	RCHY1	4	: 76,416,387	9.000 x 10-7	*	NHGRI	22491018
55	rs13031237	intron	REL	2	: 61,136,129	8.000 x 10-7	*	NHGRI	20453842
56	rs16977065	intron	RIT2	18	: 40,437,651	1.000 x 10-7	*	NHGRI	22491018
57	rs1809529	intron	SLC6A11	3	: 10,877,611	3.000 x 10-7	*	NHGRI	22491018
58	rs934734	intron	SPRED2	2	: 65,595,586	5.000 x 10-10	*	NHGRI	20453842
59	rs11121380	intron	SPSB1	1	: 9,408,959	5.000 x 10-8	*	NHGRI	22491018
60	rs7574865	intron	STAT4	2	: 191,964,633	3.000 x 10-7	*	NHGRI	20453842
61	rs7574865	intron	STAT4	2	: 191,964,633	2.000 x 10-7	*	NHGRI	20453841
62	rs3761847	intron	TRAF1	9	: 123,690,239	4.000 x 10-14	*	NHGRI	17804836
63	rs3761847	intron	TRAF1	9	: 123,690,239	2.000 x 10-7	*	NHGRI	20453842
64	rs12831974	intron	TRHDE	12	: 72,724,034	6.000 x 10-7	**2b**	NHGRI	21452313
65	rs11203203	intron	UBASH3A	21	: 43,836,186	4.000 x 10-7	*	NHGRI	20453842
66	rs12046117	intron	VTCN1	1	: 117,751,365	1.000 x 10-7	*	NHGRI	19116933
67	rs1543922	intron	ZNF175	19	: 52,084,836	3.000 x 10-7	*	NHGRI	22491018
68	rs2240335	cds-synon	PADI4	1	: 17,674,537	2.000 x 10-8	*	NHGRI	21452313
69	rs2240335	cds-synon	PADI4	1	: 17,674,537	2.000 x 10-8	*	NHGRI	21505073
70	rs11676922	intergenic	AFF3, LONRF2	2	: 100,806,940	1.000 x 10-14	*	NHGRI	20453842
71	rs10865035	intergenic	AFF3, LONRF2	2	: 100,835,734	2.000 x 10-7	*	NHGRI	20453842
72	rs1898036	intergenic	ATPBD4, COX6CP4	15	: 36,349,846	2.000 x 10-7	*	NHGRI	22491018
73	rs11900673	intergenic	B3GNT2, TMEM17	2	: 62,452,661	1.000 x 10-8	*	NHGRI	22446963
74	rs2002842	intergenic	BDP1P, SALL3	18	: 76,409,597	6.000 x 10-7	*	NHGRI	18668548
75	rs11051970	intergenic	BICD1, FGD4	12	: 32,537,488	1.000 x 10-7	*	NHGRI	22491018
76	rs12565755	intergenic	C1orf87, NFIA	1	: 61,041,875	5.000 x 10-8	*	NHGRI	22491018
77	rs874040	intergenic	C4orf52, RBPJ	4	: 26,108,197	1.000 x 10-16	**2b**	NHGRI	20453842
78	rs951005	intergenic	C9orf144B, C9orf144	9	: 34,743,681	4.000 x 10-10	*	NHGRI	20453842
79	rs11937061	intergenic	CCNG2, CXCL13	4	: 78,136,933	2.000 x 10-7	*	NHGRI	22491018
80	rs657075	intergenic	CSF2, P4HA2	5	: 131,430,118	3.000 x 10-10	*	NHGRI	22446963
81	rs6138150	intergenic	CST2, CST5	20	: 23,847,009	3.000 x 10-7	*	NHGRI	18615156
82	rs1273516	intergenic	CYP4F22, RPL23AP2	19	: 15,677,710	9.000 x 10-7	*	NHGRI	22491018
83	rs2837960	intergenic	DSCAM, C21orf130	21	: 42,511,918	2.000 x 10-7	*	NHGRI	17554300
84	rs6138892	intergenic	EBF4, RPL19P1	20	: 2,755,488	3.000 x 10-7	*	NHGRI	22491018
85	rs4937362	intergenic	ETS1, FLI1	11	: 128,492,739	8.000 x 10-7	*	NHGRI	22446963
86	rs2736340	intergenic	FAM167A, BLK	8	: 11,343,973	6.000 x 10-9	*	NHGRI	19503088
87	rs9604529	intergenic	FLJ44054	13	: 114,622,597	7.000 x 10-7	*	NHGRI	22491018
88	rs16938910	intergenic	GDAP1, PCBP2P2	8	: 75,373,948	4.000 x 10-7	*	NHGRI	22491018
89	rs12109285	intergenic	GUSBP1, CDH12	5	: 21,749,348	1.000 x 10-7	*	NHGRI	22491018
90	rs1610677	intergenic	HCP5P12, HLA-G	6	: 29,789,171	4.000 x 10-15	*	NHGRI	21653640
91	rs6457620	intergenic	HLA-DQB1, HLA-DQA	6	: 32,663,999	4.000 x 10-186	*	NHGRI	18794853
92	rs6457617	intergenic	HLA-DQB1, HLA-DQA	6	: 32,663,851	5.000 x 10-75	*	NHGRI	17554300
93	rs13192471	intergenic	HLA-DQB1, HLA-DQA	6	: 32,671,103	2.000 x 10-58	*	NHGRI	20453841
94	rs7765379	intergenic	HLA-DQB1, HLA-DQA	6	: 32,680,928	5.000 x 10-23	*	NHGRI	21452313
95	rs6457617	intergenic	HLA-DQB1, HLA-DQA	6	: 32,663,851	1.000 x 10-9	*	NHGRI	18668548
96	rs660895	intergenic	HLA-DRB1, HLA-DQA	6	: 32,577,380	1.000 x 10-108	**1f**	NHGRI	17804836
97	rs9272219	intergenic	HLA-DRB1, HLA-DQA	6	: 32,602,269	1.000 x 10-45	*	NHGRI	21653640
98	rs615672	intergenic	HLA-DRB1, HLA-DQA	6	: 32,574,171	8.000 x 10-27	*	NHGRI	17554300
99	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,429,643	5.000 x 10-109	*	NHGRI	21653640
100	rs6028945	intergenic	HSPE1P1, MAFB	20	: 38,820,805	2.000 x 10-7	*	NHGRI	18615156
101	rs743777	intergenic	IL2RB, C1QTNF6	22	: 37,551,607	1.000 x 10-7	*	NHGRI	17554300
102	rs743777	intergenic	IL2RB, C1QTNF6	22	: 37,551,607	2.000 x 10-7	*	NHGRI	21653640
103	rs10488631	intergenic	IRF5, TNPO3	7	: 128,594,183	4.000 x 10-11	**3a**	NHGRI	20453842
104	rs2280381	intergenic	IRF8, FOXF1	16	: 86,018,633	2.000 x 10-7	*	NHGRI	22446963

105	rs7155603	intergenic	JDP2, BATF	14	: 75,960,536	1.000 x 10-7	*	NHGRI	20453842
106	rs231735	intergenic	KRT18P39, CTLA4	2	: 204,693,876	6.000 x 10-9	*	NHGRI	19503088
107	rs983332	intergenic	LMO4, RPL36AP10	1	: 88,132,380	5.000 x 10-7	*	NHGRI	18615156
108	rs6774280	intergenic	MRPS35P1, MRPS36	3	: 6,255,997	9.000 x 10-7	*	NHGRI	22491018
109	rs9296015	intergenic	NOTCH4, C6orf10	6	: 32,218,989	2.000 x 10-38	*	NHGRI	21505073
110	rs1406428	intergenic	NRXN1, CRYGGP	2	: 51,736,994	2.000 x 10-7	*	NHGRI	22491018
111	rs6920220	intergenic	OLIG3, TNFAIP3	6	: 138,006,504	9.000 x 10-13	*	NHGRI	20453842
112	rs10499194	intergenic	OLIG3, TNFAIP3	6	: 138,002,637	1.000 x 10-9	*	NHGRI	17982456
113	rs6920220	intergenic	OLIG3, TNFAIP3	6	: 138,006,504	2.000 x 10-9	*	NHGRI	18794853
114	rs6920220	intergenic	OLIG3, TNFAIP3	6	: 138,006,504	1.000 x 10-7	*	NHGRI	17982456
115	rs1329568	intergenic	PAX5, RPL32P21	9	: 37,037,976	8.000 x 10-7	*	NHGRI	22491018
116	rs16906916	intergenic	PCDH15, GAPDHP21	10	: 56,848,985	8.000 x 10-7	**3b**	NHGRI	22491018
117	rs881375	intergenic	PHF19, TRAF1	9	: 123,652,898	4.000 x 10-8	**2b**	NHGRI	19503088
118	rs12131057	intergenic	POU3F1, RRAGC	1	: 38,624,129	4.000 x 10-7	*	NHGRI	20453842
119	rs10945919	intergenic	QKI, C6orf118	6	: 164,186,677	3.000 x 10-7	*	NHGRI	18615156
120	rs13017599	intergenic	REL, RPS12P3	2	: 61,164,331	2.000 x 10-12	*	NHGRI	19503088
121	rs12529514	intergenic	RNF182, CD83	6	: 14,096,658	2.000 x 10-8	*	NHGRI	22446963
122	rs437943	intergenic	RPL31P31, ARAP2	4	: 35,372,098	4.000 x 10-7	*	NHGRI	18615156
123	rs17374222	intergenic	RPLP1, GEMIN8P1	15	: 69,995,344	2.000 x 10-7	*	NHGRI	20453842
124	rs11761231	intergenic	RPS14P10, RPS15AP2	7	: 131,370,039	4.000 x 10-7	*	NHGRI	17554300
125	rs6679677	intergenic	RPS2P14, RSBN1	1	: 114,303,808	6.000 x 10-42	*	NHGRI	18794853
126	rs6679677	intergenic	RPS2P14, RSBN1	1	: 114,303,808	6.000 x 10-25	*	NHGRI	17554300
127	rs72991	intergenic	SC5DL, SORL1	11	: 121,243,716	5.000 x 10-7	*	NHGRI	22491018
128	rs7164176	intergenic	SV2B, TRNAY16P	15	: 92,211,774	5.000 x 10-7	**3a**	NHGRI	22491018
129	rs800586	intergenic	TRPS1, EIF3H	8	: 116,813,905	2.000 x 10-7	*	NHGRI	22491018
130	rs2872507	intergenic	ZPBP2, GSDMB	17	: 38,040,763	9.000 x 10-7	*	NHGRI	20453842

Ankylosing Spondylitis AS

#	rs #	Context	Gene		Location	P-value	RegulomeDB score	Source	PubMed
1	rs10781500	nearGene-5	CARD9	9	: 139,269,338	1.000 x 10-7	*	NHGRI	21743469
2	rs30187	missense	ERAP1	5	: 96,124,330	2.000 x 10-27	*	NHGRI	21743469
3	rs11209026	missense	IL23R	1	: 67,705,958	2.000 x 10-17	*	NHGRI	21743469
4	rs11209026	missense	IL23R	1	: 67,705,958	9.000 x 10-14	*	NHGRI	20062062
5	rs17095830	intron	ANO6	12	: 45,774,908	2.000 x 10-8	*	NHGRI	22138694
6	rs4389526	intron	ANTXR2	4	: 80,946,475	9.000 x 10-8	*	NHGRI	21743469
7	rs4333130	intron	ANTXR2	4	: 80,949,829	9.000 x 10-8	*	NHGRI	20062062
8	rs1326986	intron	C10orf112	10	: 19,929,513	4.000 x 10-7	*	NHGRI	20062062
9	rs2075726	intron	CSF2RB	22	: 37,310,046	9.000 x 10-7	*	NHGRI	22138694
10	rs13210693	intron	FLJ37396	6	: 109,598,964	9.000 x 10-7	*	NHGRI	22138694
11	rs2297909	intron	KIF21B	1	: 200,960,307	5.000 x 10-12	*	NHGRI	21743469
12	rs27434	cds-synon	ERAP1	5	: 96,129,512	5.000 x 10-12	*	NHGRI	20062062
13	rs10865331	intergenic	B3GNT2, TMEM17	2	: 62,551,472	7.000 x 10-34	*	NHGRI	21743469
14	rs10865331	intergenic	B3GNT2, TMEM17	2	: 62,551,472	2.000 x 10-19	*	NHGRI	20062062
15	rs10440635	intergenic	DAB2, PTGER4	5	: 40,490,790	3.000 x 10-7	*	NHGRI	21743469
16	rs7743761	intergenic	DHFRP2, HLA-S	6	: 31,336,100	5.000 x 10-304	*	NHGRI	20062062
17	rs13210693	intergenic	FLJ37396, CCDC162	6	: 109,598,964	9.000 x 10-7	*	NHGRI	22138694
18	rs2242944	intergenic	FLJ45139, RPL23AP1	21	: 40,465,178	8.000 x 10-20	*	NHGRI	20062062
19	rs378108	intergenic	FLJ45139, RPL23AP1	21	: 40,469,520	2.000 x 10-11	*	NHGRI	21743469
20	rs4349859	intergenic	HLA-S, MICA	6	: 31,365,787	1.000 x 10-200	*	NHGRI	21743469
21	rs6556416	intergenic	IL12B, ADRA1B	5	: 158,818,745	2.000 x 10-8	*	NHGRI	21743469
22	rs2310173	intergenic	IL1R2, IL1R1	2	: 102,663,628	5.000 x 10-7	*	NHGRI	20062062
23	rs8070463	intergenic	KPNB1, TBKBP1	17	: 45,768,836	5.000 x 10-8	*	NHGRI	21743469
24	rs11616188	intergenic	LTBR, RPL31P10	12	: 6,502,742	4.000 x 10-12	*	NHGRI	21743469
25	rs12146962	intergenic	MTCO1P2, NPAS3	14	: 33,381,098	9.000 x 10-7	*	NHGRI	22138694
26	rs4552569	intergenic	RPL13AP14, EDIL3	5	: 83,173,593	9.000 x 10-10	*	NHGRI	22138694
27	rs11249215	intergenic	RUNX3, SYF2	1	: 25,297,184	9.000 x 10-11	*	NHGRI	21743469
28	rs1018326	intergenic	UBE2E3, ITGA4	2	: 182,007,800	2.000 x 10-7	*	NHGRI	20062062

*

Crohn Disease, CD

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	PubMed
1	rs504963	UTR-3	FUT2	19 : 49,208,865	2.000 x 10-8	*	NHGRI	20570966
2	rs10210302	nearGene-5	ATG16L1	2 : 234,158,839	5.000 x 10-14	*	NHGRI	17554300
3	rs12677663	nearGene-5	C8orf84	8 : 74,007,347	2.000 x 10-8	*	NHGRI	22412388
4	rs11190140	nearGene-5	NKX2-3	10 : 101,291,593	3.000 x 10-16	*	NHGRI	18587394
5	rs11190141	nearGene-5	NKX2-3	10 : 101,292,390	5.000 x 10-7	*	NHGRI	22412388
6	rs11574514	nearGene-5	PSMB10	16 : 67,971,380	2.000 x 10-7	*	NHGRI	22412388
7	rs2241880	missense	ATG16L1	2 : 234,183,368	1.000 x 10-13	*	NHGRI	17435756
8	rs2241880	missense	ATG16L1	2 : 234,183,368	1.000 x 10-12	*	NHGRI	22412388
9	rs2241880	missense	ATG16L1	2 : 234,183,368	3.000 x 10-7	*	NHGRI	20570966
10	rs3764147	missense	C13orf31 (LACC1)	13 : 44,457,925	2.000 x 10-13	**1f**	NHGRI	18587394
11	rs4077515	missense	CARD9	9 : 139,266,496	1.000 x 10-36	**1f**	NHGRI	21102463
12	rs11209026	missense	IL23R	1 : 67,705,958	1.000 x 10-64	*	NHGRI	21102463
13	rs11209026	missense	IL23R	1 : 67,705,958	4.000 x 10-21	*	NHGRI	22293688
14	rs11209026	missense	IL23R	1 : 67,705,958	1.000 x 10-18	*	NHGRI	22412388
15	rs11209026	missense	IL23R	1 : 67,705,958	2.000 x 10-18	*	NHGRI	17447842
16	rs3197999	missense	MST1	3 : 49,721,532	6.000 x 10-17	*	NHGRI	21102463
17	rs3197999	missense	MST1	3 : 49,721,532	1.000 x 10-12	*	NHGRI	18587394
18	rs2476601	missense	PTPN22	1 : 114,377,568	1.000 x 10-8	**2b**	NHGRI	18587394
19	rs12720356	missense	TYK2	19 : 10,469,975	1.000 x 10-12	*	NHGRI	21102463
20	rs3792109	intron	ATG16L1	2 : 234,184,417	7.000 x 10-41	no data	NHGRI	21102463
21	rs3828309	intron	ATG16L1	2 : 234,180,410	2.000 x 10-32	**4**	NHGRI	18587394
22	rs1847472	intron	BACH2	6 : 90,973,159	5.000 x 10-9	*	NHGRI	21102463
23	rs102275	intron	C11orf10	11 : 61,557,803	2.000 x 10-11	*	NHGRI	21102463
24	rs12521868	intron	C5orf56	5 : 131,784,393	1.000 x 10-20	*	NHGRI	21102463
25	rs2188962	intron	C5orf56	5 : 131,770,805	2.000 x 10-18	**2a**	NHGRI	18587394
26	rs2188962	intron	C5orf56	5 : 131,770,805	1.000 x 10-7	*	NHGRI	20570966
27	rs6908425	intron	CDKAL1	6 : 20,728,731	9.000 x 10-10	*	NHGRI	18587394
28	rs151181	intron	CLN3	16 : 28,490,517	2.000 x 10-11	*	NHGRI	21102463
29	rs1998598	intron	DENND1B	1 : 197,727,642	9.000 x 10-9	*	NHGRI	21102463
30	rs13428812	intron	DNMT3A	2 : 25,492,467	9.000 x 10-10	*	NHGRI	21102463
31	rs2549794	intron	ERAP2	5 : 96,244,549	1.000 x 10-10	**1f**	NHGRI	21102463
32	rs2301436	intron	FGFR1OP	6 : 167,437,988	1.000 x 10-12	*	NHGRI	18587394
33	rs2301436	intron	FGFR1OP	6 : 167,437,988	6.000 x 10-8	*	NHGRI	20570966
34	rs780093	intron	GCKR	2 : 27,742,603	5.000 x 10-11	*	NHGRI	21102463
35	rs8005161	intron	GPR65	14 : 88,472,595	4.000 x 10-18	*	NHGRI	21102463
36	rs2058660	intron	IL18RAP	2 : 103,054,449	2.000 x 10-12	*	NHGRI	21102463
37	rs11465804	intron	IL23R	1 : 67,702,526	7.000 x 10-63	*	NHGRI	18587394
38	rs7517847	intron	IL23R	1 : 67,681,669	3.000 x 10-12	*	NHGRI	17435756
39	rs11805303	intron	IL23R	1 : 67,675,516	6.000 x 10-12	*	NHGRI	17554300
40	rs11465804	intron	IL23R	1 : 67,702,526	1.000 x 10-6	*	NHGRI	20570966
41	rs12722489	intron	IL2RA	10 : 6,102,012	3.000 x 10-9	*	NHGRI	21102463
42	rs2274910	intron	ITLN1	1 : 160,852,046	1.000 x 10-9	*	NHGRI	18587394
43	rs1793004	intron	NELL1	11 : 20,698,929	3.000 x 10-6	*	NHGRI	17684544
44	rs2076756	intron	NOD2	16 : 50,756,881	4.000 x 10-69	**5**	NHGRI	21102463
45	rs2076756	intron	NOD2	16 : 50,756,881	1.000 x 10-37	*	NHGRI	22412388
46	rs2076756	intron	NOD2	16 : 50,756,881	1.000 x 10-21	*	NHGRI	17684544
47	rs5743289	intron	NOD2	16 : 50,756,774	6.000 x 10-17	*	NHGRI	17804789
48	rs2076756	intron	NOD2	16 : 50,756,881	7.000 x 10-14	*	NHGRI	17435756
49	rs17221417	intron	NOD2	16 : 50,739,582	4.000 x 10-11	*	NHGRI	17554300
50	rs5743289	intron	NOD2	16 : 50,756,774	1.000 x 10-7	*	NHGRI	17447842
51	rs2797685	intron	PER3	1 : 7,879,063	7.000 x 10-9	*	NHGRI	21102463
52	rs6738825	intron	PLCL1	2 : 198,896,895	4.000 x 10-9	*	NHGRI	21102463
53	rs13003464	intron	PUS10	2 : 61,186,829	5.000 x 10-9	*	NHGRI	22412388

54	rs10181042	intron	PUS10	2	: 61,224,259	7.000 x 10-9	*	NHGRI	21102463
55	rs17309827	intron	SLC22A23	6	: 3,433,318	7.000 x 10-9	*	NHGRI	21102463
56	rs17293632	intron	SMAD3	15	: 67,442,596	3.000 x 10-19	**2a**	NHGRI	21102463
57	rs7423615	intron	SP140	2	: 231,116,874	3.000 x 10-13	*	NHGRI	21102463
58	rs744166	intron	STAT3	17	: 40,514,201	7.000 x 10-12	*	NHGRI	18587394
59	rs10495903	intron	THADA	2	: 43,806,918	2.000 x 10-14	*	NHGRI	21102463
60	rs4263839	intron	TNFSF15	9	: 117,566,440	3.000 x 10-10	**2b**	NHGRI	18587394
61	rs181359	intron	UBE2L3	22	: 21,928,641	5.000 x 10-16	**1f**	NHGRI	21102463
62	rs4809330	intron	ZGPAT	20	: 62,349,586	3.000 x 10-15	*	NHGRI	21102463
63	rs1250550	intron	ZMIZ1	10	: 81,060,317	1.000 x 10-30	*	NHGRI	21102463
64	rs7076156	intron	ZNF365	10	: 64,415,184	7.000 x 10-9	*	NHGRI	22412388
65	rs2066847	frameshift	NOD2	16	: 50,763,778	3.000 x 10-24	*	NHGRI	18587394
66	rs2066847	frameshift	NOD2	16	: 50,763,778	2.000 x 10-15	*	NHGRI	20570966
67	rs9858542	cds-synon	BSN	3	: 49,701,983	4.000 x 10-8	**1f**	NHGRI	17554300
68	rs9858542	cds-synon	BSN	3	: 49,701,983	5.000 x 10-8	**1f**	NHGRI	17554261
69	rs1142287	cds-synon	SCAMP3	1	: 155,230,131	2.000 x 10-13	*	NHGRI	21102463
70	rs3810936	cds-synon	TNFSF15	9	: 117,552,885	1.000 x 10-15	*	NHGRI	21102463
71	rs6545946	intergenic	B3GNT2, TMEM17	2	: 62,713,533	7.000 x 10-9	*	NHGRI	22412388
72	rs359457	intergenic	BOD1, CPEB4	5	: 173,279,842	3.000 x 10-12	*	NHGRI	21102463
73	rs1398024	intergenic	C10orf67, OTUD1	10	: 23,665,438	4.000 x 10-7	*	NHGRI	18723019
74	rs7927894	intergenic	C11orf30, LRRC32	11	: 76,301,316	1.000 x 10-9	*	NHGRI	18587394
75	rs11584383	intergenic	C1orf81, KIF21B	1	: 200,935,866	1.000 x 10-11	*	NHGRI	18587394
76	rs762421	intergenic	C21orf33, ICOSLG	21	: 45,615,561	1.000 x 10-9	*	NHGRI	18587394
77	rs13361189	intergenic	C5orf62, IRGM	5	: 150,223,387	2.000 x 10-10	*	NHGRI	17554261
78	rs1456893	intergenic	C7orf72, IKZF1	7	: 50,269,672	5.000 x 10-9	*	NHGRI	18587394
79	rs3091315	intergenic	CCL2, CCL7	17	: 32,593,665	2.000 x 10-13	*	NHGRI	21102463
80	rs3091316	intergenic	CCL2, CCL7	17	: 32,593,974	4.000 x 10-8	*	NHGRI	22412388
81	rs7807268	intergenic	CNTNAP2, RPL32P17	7	: 148,258,048	4.000 x 10-7	*	NHGRI	17554300
82	rs11742570	intergenic	DAB2, PTGER4	5	: 40,410,584	7.000 x 10-36	*	NHGRI	21102463
83	rs4613763	intergenic	DAB2, PTGER4	5	: 40,392,728	7.000 x 10-27	*	NHGRI	18587394
84	rs9292777	intergenic	DAB2, PTGER4	5	: 40,437,948	3.000 x 10-18	*	NHGRI	17554261
85	rs17234657	intergenic	DAB2, PTGER4	5	: 40,401,509	2.000 x 10-12	*	NHGRI	17554300
86	rs1373692	intergenic	DAB2, PTGER4	5	: 40,431,183	2.000 x 10-12	*	NHGRI	17447842
87	rs9292777	intergenic	DAB2, PTGER4	5	: 40,437,948	2.000 x 10-11	*	NHGRI	22412388
88	rs1992660	intergenic	DAB2, PTGER4	5	: 40,415,067	4.000 x 10-7	*	NHGRI	17684544
89	rs2062305	intergenic	FABP3P2, TNFSF11	13	: 43,052,880	5.000 x 10-10	*	NHGRI	21102463
90	rs10801047	intergenic	FAM5C, RGS18	1	: 191,559,356	3.000 x 10-8	*	NHGRI	17554261
91	rs9286879	intergenic	FASLG, TNFSF18	1	: 172,862,234	2.000 x 10-9	*	NHGRI	18587394
92	rs12035082	intergenic	FASLG, TNFSF18	1	: 172,898,377	2.000 x 10-7	*	NHGRI	17554261
93	rs2836754	intergenic	FLJ45139, RPL23AP1	21	: 40,291,740	5.000 x 10-7	*	NHGRI	17554261
94	rs281379	intergenic	FUT2, MAMSTR	19	: 49,214,274	7.000 x 10-12	*	NHGRI	21102463
95	rs4409764	intergenic	GOT1, NKX2-3	10	: 101,284,237	2.000 x 10-20	**3a**	NHGRI	21102463
96	rs10883365	intergenic	GOT1, NKX2-3	10	: 101,287,764	4.000 x 10-10	*	NHGRI	17554261
97	rs10883365	intergenic	GOT1, NKX2-3	10	: 101,287,764	6.000 x 10-8	*	NHGRI	17554300
98	rs9469220	intergenic	HLA-DQB1, HLA-DQA	6	: 32,658,310	2.000 x 10-6	**1f**	NHGRI	17554300
99	rs9258260	intergenic	IFITM4P, 3.8-1.5	6	: 29,723,161	2.000 x 10-10	**1f**	NHGRI	22412388
100	rs10045431	intergenic	IL12B, ADRA1B	5	: 158,814,533	4.000 x 10-13	*	NHGRI	18587394
101	rs10045431	intergenic	IL12B, ADRA1B	5	: 158,814,533	7.000 x 10-8	*	NHGRI	20570966
102	rs6887695	intergenic	IL12B, ADRA1B	5	: 158,822,645	9.000 x 10-7	*	NHGRI	17554261
103	rs3091338	intergenic	IL3, CSF2	5	: 131,402,738	4.000 x 10-8	*	NHGRI	22412388
104	rs7714584	intergenic	IRGM, ZNF300	5	: 150,270,420	8.000 x 10-19	*	NHGRI	21102463
105	rs11747270	intergenic	IRGM, ZNF300	5	: 150,258,867	3.000 x 10-16	*	NHGRI	18587394
106	rs1000113	intergenic	IRGM, ZNF300	5	: 150,240,076	3.000 x 10-7	*	NHGRI	17554300
107	rs6601764	intergenic	KLF6, AKR1E2	10	: 3,862,542	9.000 x 10-7	*	NHGRI	17554300
108	rs11167764	intergenic	MRPL11P2, NDFIP1	5	: 141,479,065	2.000 x 10-9	*	NHGRI	21102463
109	rs1819658	intergenic	MRPS35P3, IPMK	10	: 59,913,151	9.000 x 10-17	*	NHGRI	21102463
110	rs1736135	intergenic	NRIP1, CYCSP42	21	: 16,805,220	7.000 x 10-9	*	NHGRI	18587394
111	rs2413583	intergenic	PDGFB, RPL3	22	: 39,659,773	1.000 x 10-26	*	NHGRI	21102463
112	rs694739	intergenic	PRDX5, CCDC88B	11	: 64,097,233	6.000 x 10-10	**1f**	NHGRI	21102463

113	rs2542151	intergenic	PSMG2, PTPN2	18	: 12,779,947	5.000 x 10-17	*	NHGRI	18587394
114	rs2542151	intergenic	PSMG2, PTPN2	18	: 12,779,947	3.000 x 10-8	*	NHGRI	17554261
115	rs2542151	intergenic	PSMG2, PTPN2	18	: 12,779,947	2.000 x 10-7	*	NHGRI	17554300
116	rs6651252	intergenic	PVT1, GSDMC	8	: 129,567,181	4.000 x 10-18	*	NHGRI	21102463
117	rs10758669	intergenic	RCL1, JAK2	9	: 4,981,602	3.000 x 10-9	*	NHGRI	18587394
118	rs4902642	intergenic	RPL12P7, ZFP36L1	14	: 69,210,199	2.000 x 10-10	*	NHGRI	21102463
119	rs11175593	intergenic	RPL30P13, LRRK2	12	: 40,601,940	3.000 x 10-10	*	NHGRI	18587394
120	rs7746082	intergenic	RPL35P3, PRDM1	6	: 106,435,269	2.000 x 10-10	*	NHGRI	18587394
121	rs17582416	intergenic	RPS12P16, CUL2	10	: 35,287,650	2.000 x 10-9	*	NHGRI	18587394
122	rs3024505	intergenic	RPS14P1, IL10	1	: 206,939,904	2.000 x 10-14	**2b**	NHGRI	21102463
123	rs713875	intergenic	RPS3AP51, LIF	22	: 30,592,487	7.000 x 10-12	**1f**	NHGRI	21102463
124	rs13073817	intergenic	SATB1, KCNH8	3	: 18,706,858	7.000 x 10-9	*	NHGRI	21102463
125	rs6596075	intergenic	SLC22A5, C5orf56	5	: 131,742,228	3.000 x 10-7	**1f**	NHGRI	17554300
126	rs11229030	intergenic	SLC43A3, RTN4RL2	11	: 57,203,009	8.000 x 10-9	*	NHGRI	22412388
127	rs736289	intergenic	SLC7A10, CEBPA	19	: 33,757,062	9.000 x 10-9	*	NHGRI	21102463
128	rs7705924	intergenic	SLCO6A1, PAM	5	: 101,946,798	2.000 x 10-8	*	NHGRI	22412388
129	rs212388	intergenic	TAGAP, FNDC1	6	: 159,490,436	2.000 x 10-11	*	NHGRI	21102463
130	rs10734105	intergenic	TCERG1L, FLJ46300	10	: 133,172,119	3.000 x 10-8	*	NHGRI	22412388
131	rs7702331	intergenic	TMEM174, FOXD1	5	: 72,551,134	6.000 x 10-12	*	NHGRI	21102463
132	rs1551398	intergenic	TRIB1, FAM84B	8	: 126,540,051	5.000 x 10-9	*	NHGRI	18587394
133	rs1906493	intergenic	TRIB1, FAM84B	8	: 127,092,882	3.000 x 10-7	*	NHGRI	22412388
134	rs10761659	intergenic	ZNF365, ALDH7A1P4	10	: 64,445,564	4.000 x 10-22	*	NHGRI	21102463
135	rs10995271	intergenic	ZNF365, ALDH7A1P4	10	: 64,438,486	4.000 x 10-20	*	NHGRI	18587394
136	rs224136	intergenic	ZNF365, ALDH7A1P4	10	: 64,470,675	1.000 x 10-10	*	NHGRI	17435756
137	rs10761659	intergenic	ZNF365, ALDH7A1P4	10	: 64,445,564	2.000 x 10-7	*	NHGRI	17554300
138	rs2872507	intergenic	ZPBP2, GSDMB	17	: 38,040,763	5.000 x 10-9	*	NHGRI	18587394

Inflammatory Bowel Diseases, IBD

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	Study
1	rs10889677	UTR-3	IL23R	1 : 67,725,120	9.037 x 10-11	*	dbGaP	phs000130
2	rs11209026	missense	IL23R	1 : 67,705,958	4.000 x 10-11	*	NHGRI	
3	rs11209026	missense	IL23R	1 : 67,705,958	4.592 x 10-11	*	dbGaP	phs000130
4	rs11209026	missense	IL23R	1 : 67,705,958	7.000 x 10-11	*	NHGRI	
5	rs2315008	intron	ZGPAT	20 : 62,343,956	9.000 x 10-15	*	NHGRI	
6	rs2076756	intron	NOD2	16 : 50,756,881	1.262 x 10-14	*	dbGaP	phs000130
7	rs7517847	intron	IL23R	1 : 67,681,669	2.991 x 10-13	*	dbGaP	phs000130
8	rs7517847	intron	IL23R	1 : 67,681,669	4.000 x 10-13	*	NHGRI	
9	rs1343151	intron	IL23R	1 : 67,719,129	1.628 x 10-11	*	dbGaP	phs000130
10	rs10489629	intron	IL23R	1 : 67,688,349	6.790 x 10-11	*	dbGaP	phs000130
11	rs2201841	intron	IL23R	1 : 67,694,202	3.574 x 10-10	*	dbGaP	phs000130
12	rs11465804	intron	IL23R	1 : 67,702,526	3.737 x 10-10	*	dbGaP	phs000130
13	rs5743289	intron	NOD2	16 : 50,756,774	4.000 x 10-10	*	NHGRI	
14	rs2076756	intron	NOD2	16 : 50,756,881	5.000 x 10-10	*	NHGRI	
15	rs1004819	intron	IL23R	1 : 67,670,213	1.504 x 10-9	*	dbGaP	phs000130
16	rs8049439	intron	ATXN2L	16 : 28,837,515	2.000 x 10-9	**1b**	NHGRI	
17	rs2412973	intron	HORMAD2	22 : 30,529,631	2.000 x 10-9	*	NHGRI	
18	rs1250550	intron	ZMIZ1	10 : 81,060,317	6.000 x 10-9	*	NHGRI	
19	rs2066843	cds-synon	NOD2	16 : 50,745,199	7.869 x 10-13	*	dbGaP	phs000130
20	rs9271366	intergenic	HLA-DRB1, HLA-DQA	6 : 32,586,854	2.000 x 10-70	*	NHGRI	
21	rs9271366	intergenic	HLA-DRB1, HLA-DQA	6 : 32,586,854	3.000 x 10-31	*	NHGRI	
22	rs2006996	intergenic	TNFSF15, TNFSF8	9 : 117,592,638	4.000 x 10-16	*	NHGRI	
23	rs2006996	intergenic	TNFSF15, TNFSF8	9 : 117,592,638	4.000 x 10-13	*	NHGRI	
24	rs2836878	intergenic	FLJ45139, RPL23AP1	21 : 40,465,534	4.000 x 10-12	*	NHGRI	
25	rs9271366	intergenic	HLA-DRB1, HLA-DQA	6 : 32,586,854	8.000 x 10-11	*	NHGRI	
26	rs10500264	intergenic	SLC7A10, CEBPA	19 : 33,750,314	4.000 x 10-10	*	NHGRI	
27	rs11209032	intergenic	IL23R, IL12RB2	1 : 67,740,092	8.645 x 10-10	*	dbGaP	phs000130
28	rs477515	intergenic	HLA-DRB1, HLA-DQA	6 : 32,569,691	1.000 x 10-8	*	NHGRI	

Psoriasis, PS

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	Study
1	rs8192583	UTR-5	GPSM3	6 : 32,163,274	2.250 x 10-20	**1f**	dbGaP	phs000019
2	rs3094187	UTR-5	TCF19	6 : 31,126,944	3.725 x 10-13	**1f**	dbGaP	phs000019
3	rs2240803	UTR-3	DPCR1	6 : 30,920,957	8.193 x 10-13	*	dbGaP	phs000019
4	rs2395029	UTR-3	HCP5	6 : 31,431,780	2.000 x 10-26	*	NHGRI	
5	rs8365	UTR-3	RNF5	6 : 32,148,403	1.094 x 10-17	**1f**	dbGaP	phs000019
6	rs3130453	STOP-GAIN	CCHCR1	6 : 31,124,849	2.057 x 10-12	*	dbGaP	phs000019
7	rs3132965	nearGene-5	AGPAT1	6 : 32,146,997	8.389 x 10-16	*	dbGaP	phs000019
8	rs3094187	nearGene-5	CCHCR1	6 : 31,126,944	3.725 x 10-13	*	dbGaP	phs000019
9	rs7773175	nearGene-5	HLA-C	6 : 31,240,959	4.772 x 10-30	*	dbGaP	phs000019
10	rs2249742	nearGene-5	HLA-C	6 : 31,240,721	6.346 x 10-18	**1f**	dbGaP	phs000019
11	rs2395471	nearGene-5	HLA-C	6 : 31,240,692	8.836 x 10-18	**1f**	dbGaP	phs000019
12	rs2249741	nearGene-5	HLA-C	6 : 31,240,712	3.340 x 10-12	*	dbGaP	phs000019
13	rs2524082	nearGene-5	HLA-C	6 : 31,241,761	2.215 x 10-11	*	dbGaP	phs000019
14	rs9267502	nearGene-5	LST1	6 : 31,553,194	3.786 x 10-19	*	dbGaP	phs000019
15	rs2734573	nearGene-5	MCCD1	6 : 31,494,738	1.499 x 10-12	*	dbGaP	phs000019
16	rs176095	nearGene-5	PBX2	6 : 32,158,319	2.183 x 10-12	**1f**	dbGaP	phs000019
17	rs3130453	nearGene-5	TCF19	6 : 31,124,849	2.057 x 10-12	*	dbGaP	phs000019
18	rs2021723	nearGene-5	TRIM40	6 : 30,103,923	8.690 x 10-15	*	dbGaP	phs000019
19	rs8365	nearGene-3	AGER	6 : 32,148,403	1.094 x 10-17	*	dbGaP	phs000019
20	rs1265086	nearGene-3	CCHCR1	6 : 31,109,882	2.689 x 10-11	**1f**	dbGaP	phs000019
21	rs9468843	nearGene-3	DDR1	6 : 30,867,958	9.009 x 10-14	*	dbGaP	phs000019
22	rs176095	nearGene-3	GPSM3	6 : 32,158,319	2.183 x 10-12	*	dbGaP	phs000019
23	rs2853950	nearGene-3	HLA-C	6 : 31,236,175	1.756 x 10-14	*	dbGaP	phs000019
24	rs11575907	missense	HLA-DOB	6 : 32,782,112	5.960 x 10-19	*	dbGaP	phs000019
25	rs20541	missense	IL13	5 : 131,995,964	5.000 x 10-15	*	NHGRI	
26	rs11209026	missense	IL23R	1 : 67,705,958	7.000 x 10-7	*	NHGRI	
27	rs8192591	missense	NOTCH4	6 : 32,185,796	3.403 x 10-24	*	dbGaP	phs000019
28	rs33980500	missense	TRAF3IP2	6 : 111,913,262	1.000 x 10-16	*	NHGRI	
29	rs12720356	missense	TYK2	19 : 10,469,975	4.000 x 10-11	*	NHGRI	
30	rs2295663	intron	BAT5	6 : 31,669,295	8.396 x 10-24	*	dbGaP	phs000019
31	rs9267673	intron	C2	6 : 31,883,679	2.300 x 10-33	**1f**	dbGaP	phs000019
32	rs6906662	intron	C6orf10	6 : 32,266,506	2.109 x 10-14	*	dbGaP	phs000019
33	rs27524	intron	CAST	5 : 96,101,944	3.000 x 10-11	*	NHGRI	
34	rs1265078	intron	CCHCR1	6 : 31,112,602	2.039 x 10-21	*	dbGaP	phs000019
35	rs7993214	intron	COG6	13 : 40,350,912	2.000 x 10-6	*	NHGRI	
36	rs2239518	intron	DDR1	6 : 30,865,725	3.454 x 10-12	**2b**	dbGaP	phs000019
37	rs27524	intron	ERAP1	5 : 96,101,944	3.000 x 10-11	*	NHGRI	
38	rs10782001	intron	FBXL19	16 : 30,942,625	9.000 x 10-10	*	NHGRI	
39	rs3213094	intron	IL12B	5 : 158,750,769	3.000 x 10-26	*	NHGRI	
40	rs3213094	intron	IL12B	5 : 158,750,769	5.000 x 10-11	*	NHGRI	
41	rs2201841	intron	IL23R	1 : 67,694,202	3.000 x 10-8	*	NHGRI	
42	rs17716942	intron	KCNH7	2 : 163,260,691	1.000 x 10-13	*	NHGRI	
43	rs12586317	intron	KIAA0391	14 : 35,682,172	2.000 x 10-8	*	NHGRI	
44	rs9295938	intron	MUC21	6 : 30,953,105	9.163 x 10-18	*	dbGaP	phs000019
45	rs4795067	intron	NOS2	17 : 26,106,675	4.000 x 10-11	*	NHGRI	
46	rs3823418	intron	PSORS1C1	6 : 31,100,942	1.169 x 10-33	*	dbGaP	phs000019
47	rs3094205	intron	PSORS1C1	6 : 31,091,862	1.733 x 10-15	*	dbGaP	phs000019
48	rs3130573	intron	PSORS1C1	6 : 31,106,268	6.212 x 10-15	**1f**	dbGaP	phs000019
49	rs3130573	intron	PSORS1C2	6 : 31,106,268	6.212 x 10-15	*	dbGaP	phs000019
50	rs240993	intron	REV3L	6 : 111,673,714	5.000 x 10-20	*	NHGRI	
51	rs3132965	intron	RNF5	6 : 32,146,997	8.389 x 10-16	*	dbGaP	phs000019
52	rs2066808	intron	STAT2	12 : 56,737,973	1.000 x 10-9	*	NHGRI	
53	rs2066808	intron	STAT2	12 : 56,737,973	2.000 x 10-7	*	NHGRI	
54	rs3093662	intron	TNF	6 : 31,544,189	4.752 x 10-21	*	dbGaP	phs000019

55	rs610604	intron	TNFAIP3	6	: 138,199,417	9.000 x 10-12	*	NHGRI	
56	rs610604	intron	TNFAIP3	6	: 138,199,417	7.000 x 10-7	*	NHGRI	
57	rs2077580	intron	TNXB	6	: 32,020,844	7.445 x 10-32	*	dbGaP	phs000019
58	rs2107195	intron	TRIM15	6	: 30,137,866	1.675 x 10-17	*	dbGaP	phs000019
59	rs1076160	intron	TSC1	9	: 135,776,034	6.000 x 10-7	*	NHGRI	
60	rs280519	intron	TYK2	19	: 10,472,933	4.000 x 10-9	*	NHGRI	
61	rs8192583	intron	NOTCH4	6	: 32,163,274	2.250 x 10-20	*	dbGaP	phs000019
62	rs443198	cds-synon	NOTCH4	6	: 32,190,406	1.526 x 10-11	**1f**	dbGaP	phs000019
63	rs495337	cds-synon	SPATA2	20	: 48,522,330	1.000 x 10-8	*	NHGRI	
64	rs495337	cds-synon	SPATA2	20	: 48,522,330	2.000 x 10-7	*	NHGRI	
65	rs7762370	intergenic	BTNL2, HLA-DRA	6	: 32,400,190	5.130 x 10-19	*	dbGaP	phs000019
66	rs9501624	intergenic	BTNL2, HLA-DRA	6	: 32,399,286	2.357 x 10-12	*	dbGaP	phs000019
67	rs1975974	intergenic	C17orf51, UBBP4	17	: 21,707,060	1.000 x 10-7	*	NHGRI	
68	rs3130955	intergenic	HCG22, C6orf15	6	: 31,054,511	1.312 x 10-20	**1f**	dbGaP	phs000019
69	rs2844627	intergenic	HCG27, HLA-C	6	: 31,229,462	2.307 x 10-36	*	dbGaP	phs000019
70	rs2394895	intergenic	HCG27, HLA-C	6	: 31,206,979	2.593 x 10-29	*	dbGaP	phs000019
71	rs3130517	intergenic	HCG27, HLA-C	6	: 31,190,303	1.430 x 10-28	*	dbGaP	phs000019
72	rs3130713	intergenic	HCG27, HLA-C	6	: 31,205,617	9.808 x 10-28	*	dbGaP	phs000019
73	rs3130467	intergenic	HCG27, HLA-C	6	: 31,187,075	1.339 x 10-27	*	dbGaP	phs000019
74	rs9263967	intergenic	HCG27, HLA-C	6	: 31,186,245	7.296 x 10-15	*	dbGaP	phs000019
75	rs3130685	intergenic	HCG27, HLA-C	6	: 31,206,206	8.583 x 10-14	*	dbGaP	phs000019
76	rs3130425	intergenic	HCG27, HLA-C	6	: 31,218,327	1.808 x 10-12	*	dbGaP	phs000019
77	rs3095250	intergenic	HCG27, HLA-C	6	: 31,208,340	2.892 x 10-12	*	dbGaP	phs000019
78	rs3132496	intergenic	HCG27, HLA-C	6	: 31,208,610	2.898 x 10-12	*	dbGaP	phs000019
79	rs3132486	intergenic	HLA-C, RPL3P2	6	: 31,243,170	6.660 x 10-13	*	dbGaP	phs000019
80	rs3132485	intergenic	HLA-C, RPL3P2	6	: 31,243,389	2.623 x 10-12	*	dbGaP	phs000019
81	rs2647087	intergenic	HLA-DQB1, HLA-DQA	6	: 32,681,049	1.865 x 10-16	*	dbGaP	phs000019
82	rs2858333	intergenic	HLA-DQB1, HLA-DQA	6	: 32,681,085	8.551 x 10-16	*	dbGaP	phs000019
83	rs2856726	intergenic	HLA-DQB1, HLA-DQA	6	: 32,666,721	5.086 x 10-12	*	dbGaP	phs000019
84	rs12203586	intergenic	HLA-DQB1, HLA-DQA	6	: 32,679,591	5.743 x 10-12	**1f**	dbGaP	phs000019
85	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,429,643	8.870 x 10-13	*	dbGaP	phs000019
86	rs10484552	intergenic	HLA-E, GNL1	6	: 30,484,036	7.487 x 10-19	*	dbGaP	phs000019
87	rs13437088	intergenic	HLA-S, MICA	6	: 31,355,119	2.539 x 10-13	*	dbGaP	phs000019
88	rs10947208	intergenic	HLA-S, MICA	6	: 31,361,837	2.182 x 10-12	*	dbGaP	phs000019
89	rs7756521	intergenic	IER3, DDR1	6	: 30,848,253	7.581 x 10-19	*	dbGaP	phs000019
90	rs4711229	intergenic	IER3, DDR1	6	: 30,756,744	5.431 x 10-14	*	dbGaP	phs000019
91	rs9295917	intergenic	IER3, DDR1	6	: 30,769,772	5.091 x 10-13	*	dbGaP	phs000019
92	rs4713380	intergenic	IER3, DDR1	6	: 30,785,273	6.508 x 10-13	**2b**	dbGaP	phs000019
93	rs12526186	intergenic	IER3, DDR1	6	: 30,736,151	1.272 x 10-12	*	dbGaP	phs000019
94	rs13198118	intergenic	IER3, DDR1	6	: 30,770,732	2.074 x 10-12	**2b**	dbGaP	phs000019
95	rs7749924	intergenic	IER3, DDR1	6	: 30,797,991	2.249 x 10-12	*	dbGaP	phs000019
96	rs9295924	intergenic	IER3, DDR1	6	: 30,782,361	2.685 x 10-12	*	dbGaP	phs000019
97	rs3131043	intergenic	IER3, DDR1	6	: 30,758,466	1.814 x 10-11	**1f**	dbGaP	phs000019
98	rs2546890	intergenic	IL12B, ADRA1B	5	: 158,759,900	1.000 x 10-20	*	NHGRI	
99	rs4649203	intergenic	IL28RA, GRHL3	1	: 24,519,920	7.000 x 10-8	*	NHGRI	
100	rs4085613	intergenic	LCE3E, LCE3D	1	: 152,550,018	7.000 x 10-30	*	NHGRI	
101	rs4112788	intergenic	LCE3E, LCE3D	1	: 152,551,276	3.000 x 10-10	*	NHGRI	
102	rs9501106	intergenic	MICA, HLA-X	6	: 31,388,109	3.527 x 10-16	*	dbGaP	phs000019
103	rs9295993	intergenic	MICA, HLA-X	6	: 31,388,595	1.344 x 10-15	*	dbGaP	phs000019
104	rs9266844	intergenic	MICA, HLA-X	6	: 31,384,331	1.022 x 10-11	*	dbGaP	phs000019
105	rs7772549	intergenic	MICA, HLA-X	6	: 31,407,643	1.032 x 10-11	*	dbGaP	phs000019
106	rs9266846	intergenic	MICA, HLA-X	6	: 31,384,889	2.441 x 10-11	*	dbGaP	phs000019
107	rs17476793	intergenic	MICC, SUCLA2P1	6	: 30,410,988	1.261 x 10-17	*	dbGaP	phs000019
108	rs13191258	intergenic	MUC21, HCG22	6	: 30,978,717	3.951 x 10-33	*	dbGaP	phs000019
109	rs2844645	intergenic	MUC21, HCG22	6	: 31,015,182	7.292 x 10-19	*	dbGaP	phs000019
110	rs9366764	intergenic	MUC21, HCG22	6	: 30,979,793	8.095 x 10-18	*	dbGaP	phs000019
111	rs9394031	intergenic	MUC21, HCG22	6	: 30,991,643	8.254 x 10-18	*	dbGaP	phs000019
112	rs3871466	intergenic	MUC21, HCG22	6	: 30,983,683	7.904 x 10-17	*	dbGaP	phs000019
113	rs2523870	intergenic	MUC21, HCG22	6	: 31,014,116	2.725 x 10-12	*	dbGaP	phs000019

114	rs2517552	intergenic	MUC21, HCG22	6 : 31,007,590	8.159 x 10-12	*	dbGaP	phs000019
115	rs2894176	intergenic	MUC21, HCG22	6 : 30,986,038	9.713 x 10-12	*	dbGaP	phs000019
116	rs2517527	intergenic	MUC21, HCG22	6 : 31,021,547	1.831 x 10-11	*	dbGaP	phs000019
117	rs9262492	intergenic	MUC21, HCG22	6 : 30,986,015	2.993 x 10-11	**2b**	dbGaP	phs000019
118	rs6916062	intergenic	NOTCH4, C6orf10	6 : 32,219,041	2.524 x 10-22	*	dbGaP	phs000019
119	rs9267463	intergenic	PPIAP9, RPL15P4	6 : 31,490,480	1.104 x 10-23	*	dbGaP	phs000019
120	rs9267464	intergenic	PPIAP9, RPL15P4	6 : 31,490,646	1.806 x 10-22	*	dbGaP	phs000019
121	rs2734573	intergenic	PPIAP9, RPL15P4	6 : 31,494,738	1.499 x 10-12	*	dbGaP	phs000019
122	rs8016947	intergenic	PSMA6, RPLP0P3	14 : 35,832,666	2.000 x 10-11	*	NHGRI	
123	rs702873	intergenic	RPL21P33, REL	2 : 61,081,542	4.000 x 10-9	*	NHGRI	
124	rs842636	intergenic	RPL21P33, REL	2 : 61,091,950	6.000 x 10-6	*	NHGRI	
125	rs12191877	intergenic	RPL3P2, WASF5P	6 : 31,252,925	1.000 x 10-100	*	NHGRI	
126	rs12191877	intergenic	RPL3P2, WASF5P	6 : 31,252,925	2.024 x 10-51	*	dbGaP	phs000019
127	rs12191877	intergenic	RPL3P2, WASF5P	6 : 31,252,925	4.000 x 10-32	*	NHGRI	
128	rs12580100	intergenic	RPS26, ERBB3	12 : 56,439,209	1.000 x 10-7	**3a**	NHGRI	
129	rs6809854	intergenic	SATB1, KCNH8	3 : 18,784,423	1.000 x 10-7	*	NHGRI	
130	rs1008953	intergenic	SDC4, SYS1	20 : 43,980,726	1.000 x 10-7	*	NHGRI	
131	rs17728338	intergenic	TNIP1, ANXA6	5 : 150,478,318	1.000 x 10-20	*	NHGRI	
132	rs1015465	intergenic	TRIM31, TRIM40	6 : 30,086,340	3.685 x 10-15	*	dbGaP	phs000019
133	rs2082412	intergenic	UBLCP1, IL12B	5 : 158,717,789	2.000 x 10-28	*	NHGRI	
134	rs10484554	intergenic	WASF5P, HLA-B	6 : 31,274,555	4.000 x 10-214	*	NHGRI	
135	rs9468933	intergenic	WASF5P, HLA-B	6 : 31,265,057	1.621 x 10-46	*	dbGaP	phs000019
136	rs10484554	intergenic	WASF5P, HLA-B	6 : 31,274,555	2.000 x 10-39	*	NHGRI	
137	rs2894207	intergenic	WASF5P, HLA-B	6 : 31,263,751	3.766 x 10-38	*	dbGaP	phs000019
138	rs9380237	intergenic	WASF5P, HLA-B	6 : 31,264,392	2.576 x 10-28	*	dbGaP	phs000019
139	rs2524163	intergenic	WASF5P, HLA-B	6 : 31,259,579	5.329 x 10-20	*	dbGaP	phs000019
140	rs2243868	intergenic	WASF5P, HLA-B	6 : 31,261,276	2.798 x 10-19	*	dbGaP	phs000019
141	rs2853923	intergenic	WASF5P, HLA-B	6 : 31,265,737	2.701 x 10-15	*	dbGaP	phs000019
142	rs9380240	intergenic	WASF5P, HLA-B	6 : 31,268,832	4.412 x 10-14	*	dbGaP	phs000019
143	rs2442719	intergenic	WASF5P, HLA-B	6 : 31,320,538	2.809 x 10-13	*	dbGaP	phs000019
144	rs3873386	intergenic	WASF5P, HLA-B	6 : 31,273,745	1.173 x 10-12	*	dbGaP	phs000019
145	rs2156875	intergenic	WASF5P, HLA-B	6 : 31,317,347	3.707 x 10-12	*	dbGaP	phs000019
146	rs9468937	intergenic	WASF5P, HLA-B	6 : 31,270,118	4.828 x 10-12	*	dbGaP	phs000019
147	rs3134792	intergenic	WASF5P, HLA-B	6 : 31,312,326	1.000 x 10-9	*	NHGRI	
148	rs9267673	intergenic	ZBTB12, C2	6 : 31,883,679	2.300 x 10-33	*	dbGaP	phs000019

Ulcerative Colitis, UC

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	PubMed
1	rs10889677	UTR-3	IL23R	1 : 67,725,120	1.000 x 10-8	**no data**	NHGRI	19122664
2	rs2297441	UTR-3	RTEL1	20 : 62,327,582	2.000 x 10-10	**4**	NHGRI	21297633
3	rs10781500	nearGene-5	CARD9	9 : 139,269,338	7.000 x 10-6	**no data**	NHGRI	19915572
4	rs678170	nearGene-5	FAM55A	11 : 114,431,956	5.000 x 10-14	**5**	NHGRI	21297633
5	rs907611	nearGene-5	LSP1	11 : 1,874,072	1.000 x 10-10	**2a**	NHGRI	21297633
6	rs11190140	nearGene-5	NKX2-3	10 : 101,291,593	1.000 x 10-8	**5**	NHGRI	20228799
7	rs3806308	nearGene-5	RNF186	1 : 20,142,866	7.000 x 10-9	**2b**	NHGRI	19122664
8	rs2297441	nearGene-5	TNFRSF6B	20 : 62,327,582	2.000 x 10-10	**4**	NHGRI	21297633
9	rs1317209	nearGene-3	RNF186	1 : 20,140,036	2.000 x 10-10	**5**	NHGRI	20228799
10	rs4077515	missense	CARD9	9 : 139,266,496	5.000 x 10-8	**1f**	NHGRI	20228799
11	rs1801274	missense	FCGR2A	1 : 161,479,745	2.000 x 10-20	*	NHGRI	21297633
12	rs1801274	missense	FCGR2A	1 : 161,479,745	2.000 x 10-12	*	NHGRI	19915573
13	rs2305480	missense	GSDMB	17 : 38,062,196	3.000 x 10-8	*	NHGRI	20228799
14	rs5771069	missense	IL17REL	22 : 50,435,480	4.000 x 10-8	*	NHGRI	20228798
15	rs5771069	missense	IL17REL	22 : 50,435,480	2.000 x 10-7	*	NHGRI	21297633
16	rs11209026	missense	IL23R	1 : 67,705,958	5.000 x 10-28	*	NHGRI	21297633
17	rs11209026	missense	IL23R	1 : 67,705,958	3.000 x 10-10	*	NHGRI	19915572
18	rs11209026	missense	IL23R	1 : 67,705,958	1.000 x 10-8	*	NHGRI	19122664
19	rs3194051	missense	IL7R	5 : 35,876,274	4.000 x 10-8	*	NHGRI	21297633

20	rs3197999	missense	MST1	3	: 49,721,532	4.000 x 10-9	*	NHGRI	20228799
21	rs17388568	intron	ADAD1	4	: 123,329,362	9.000 x 10-7	*	NHGRI	21297633
22	rs9822268	intron	APEH	3	: 49,719,729	2.000 x 10-17	*	NHGRI	21297633
23	rs7554511	intron	C1orf106	1	: 200,877,562	2.000 x 10-13	*	NHGRI	21297633
24	rs7554511	intron	C1orf106	1	: 200,877,562	1.000 x 10-7	*	NHGRI	19915572
25	rs9263739	intron	CCHCR1	6	: 31,111,356	4.000 x 10-67	*	NHGRI	19915573
26	rs12261843	intron	CCNY	10	: 35,554,054	7.000 x 10-10	*	NHGRI	21297633
27	rs4781011	intron	CIITA	16	: 10,975,311	3.000 x 10-7	**2b**	NHGRI	20228799
28	rs267939	intron	DAP	5	: 10,752,315	6.000 x 10-12	*	NHGRI	21297633
29	rs798502	intron	GNA12	7	: 2,789,880	3.000 x 10-15	**1b**	NHGRI	21297633
30	rs3024493	intron	IL10	1	: 206,943,968	1.000 x 10-12	**2b**	NHGRI	20228798
31	rs3024493	intron	IL10	1	: 206,943,968	8.000 x 10-8	**2b**	NHGRI	19915572
32	rs2201841	intron	IL23R	1	: 67,694,202	1.000 x 10-13	*	NHGRI	20228799
33	rs2870946	intron	IL26	12	: 68,596,661	5.000 x 10-7	*	NHGRI	19122664
34	rs2158836	intron	LAMB1	7	: 107,580,839	7.000 x 10-6	*	NHGRI	19122664
35	rs4654925	intron	OTUD3	1	: 20,227,723	9.000 x 10-22	*	NHGRI	20228798
36	rs35675666	intron	PARK7	1	: 8,021,973	5.000 x 10-9	*	NHGRI	21297633
37	rs7608910	intron	PUS10	2	: 61,204,856	2.000 x 10-14	*	NHGRI	21297633
38	rs13003464	intron	PUS10	2	: 61,186,829	7.000 x 10-9	*	NHGRI	20228799
39	rs1992950	intron	SATB2	2	: 200,290,359	5.000 x 10-7	*	NHGRI	20228799
40	rs4246905	intron	TNFSF15	9	: 117,553,249	6.000 x 10-12	*	NHGRI	21297633
41	rs1728785	intron	ZFP90	16	: 68,591,230	3.000 x 10-8	*	NHGRI	19915572
42	rs9858542	cds-synon	BSN	3	: 49,701,983	7.000 x 10-9	**1f**	NHGRI	19915572
43	rs9268480	cds-synon	BTNL2	6	: 32,363,844	3.000 x 10-7	*	NHGRI	19915573
44	rs10781499	cds-synon	CARD9	9	: 139,266,405	3.000 x 10-19	**1f**	NHGRI	21297633
45	rs2155219	intergenic	C11orf30, LRRC32	11	: 76,299,194	5.000 x 10-16	*	NHGRI	21297633
46	rs10800309	intergenic	C1orf192, FCGR2A	1	: 161,472,158	3.000 x 10-9	*	NHGRI	20228799
47	rs11584383	intergenic	C1orf81, KIF21B	1	: 200,935,866	2.000 x 10-7	*	NHGRI	20228799
48	rs2838519	intergenic	C21orf33, ICOSLG	21	: 45,615,023	6.000 x 10-11	*	NHGRI	21297633
49	rs941823	intergenic	COG6, FOXO1	13	: 41,013,977	4.000 x 10-12	*	NHGRI	21297633
50	rs9548988	intergenic	COG6, FOXO1	13	: 40,505,510	3.000 x 10-7	*	NHGRI	19915572
51	rs11676348	intergenic	CXCR2, CXCR1	2	: 219,010,146	1.000 x 10-10	**1f**	NHGRI	21297633
52	rs6451493	intergenic	DAB2, PTGER4	5	: 40,410,935	3.000 x 10-9	**3a**	NHGRI	21297633
53	rs2836878	intergenic	FLJ45139, RPL23AP12	21	: 40,465,534	2.000 x 10-22	*	NHGRI	21297633
54	rs6584283	intergenic	GOT1, NKX2-3	10	: 101,290,301	8.000 x 10-21	*	NHGRI	21297633
55	rs6584283	intergenic	GOT1, NKX2-3	10	: 101,290,301	2.000 x 10-7	*	NHGRI	19915572
56	rs6584283	intergenic	GOT1, NKX2-3	10	: 101,290,301	2.000 x 10-6	*	NHGRI	20228798
57	rs17085007	intergenic	GPR12, RPS20P32	13	: 27,531,267	1.000 x 10-16	*	NHGRI	21297633
58	rs17085007	intergenic	GPR12, RPS20P32	13	: 27,531,267	7.000 x 10-8	*	NHGRI	19915573
59	rs4676406	intergenic	GPR35, AQP12B	2	: 241,579,108	8.000 x 10-11	*	NHGRI	21297633
60	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,429,643	1.000 x 10-55	*	NHGRI	21297633
61	rs9268877	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,431,147	4.000 x 10-23	*	NHGRI	19915572
62	rs2395185	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,433,167	5.000 x 10-22	*	NHGRI	19915573
63	rs9268877	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,431,147	6.000 x 10-18	*	NHGRI	18836448
64	rs2395185	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,433,167	1.000 x 10-16	*	NHGRI	19122664
65	rs9268923	intergenic	HLA-DRB9, HLA-DRB5	6	: 32,432,835	4.000 x 10-15	*	NHGRI	20228798
66	rs6017342	intergenic	HNF4A, RPL37AP1	20	: 43,065,028	1.000 x 10-20	*	NHGRI	21297633
67	rs6017342	intergenic	HNF4A, RPL37AP1	20	: 43,065,028	9.000 x 10-17	*	NHGRI	19915572
68	rs6871626	intergenic	IL12B, ADRA1B	5	: 158,826,792	1.000 x 10-21	*	NHGRI	21297633
69	rs2310173	intergenic	IL1R2, IL1R1	2	: 102,663,628	3.000 x 10-12	*	NHGRI	21297633
70	rs10975003	intergenic	INSL6, INSL4	9	: 5,213,687	1.000 x 10-6	*	NHGRI	19915573
71	rs16940202	intergenic	IRF8, FOXF1	16	: 86,014,241	6.000 x 10-19	*	NHGRI	21297633
72	rs4728142	intergenic	KCP, IRF5	7	: 128,573,967	2.000 x 10-8	**1f**	NHGRI	21297633
73	rs1297265	intergenic	NRIP1, CYCSP42	21	: 16,817,051	7.000 x 10-13	**2a**	NHGRI	21297633
74	rs1736135	intergenic	NRIP1, CYCSP42	21	: 16,805,220	2.000 x 10-7	*	NHGRI	20228799
75	rs6920220	intergenic	OLIG3, TNFAIP3	6	: 138,006,504	8.000 x 10-17	*	NHGRI	21297633
76	rs254560	intergenic	PITX1, H2AFY	5	: 134,443,606	1.000 x 10-9	*	NHGRI	21297633
77	rs10758669	intergenic	RCL1, JAK2	9	: 4,981,602	2.000 x 10-25	*	NHGRI	21297633
78	rs10758669	intergenic	RCL1, JAK2	9	: 4,981,602	1.000 x 10-6	*	NHGRI	20228799

79	rs6426833	intergenic	RNF186, OTUD3	1	: 20,171,860	4.000 x 10-35	*	NHGRI	21297633
80	rs6426833	intergenic	RNF186, OTUD3	1	: 20,171,860	2.000 x 10-21	*	NHGRI	20228799
81	rs6426833	intergenic	RNF186, OTUD3	1	: 20,171,860	5.000 x 10-13	*	NHGRI	19122664
82	rs6426833	intergenic	RNF186, OTUD3	1	: 20,171,860	2.000 x 10-11	*	NHGRI	19915572
83	rs6911490	intergenic	RPL35P3, PRDM1	6	: 106,522,027	1.000 x 10-8	*	NHGRI	21297633
84	rs7134599	intergenic	RPL39P28, IFNG	12	: 68,500,075	1.000 x 10-16	*	NHGRI	21297633
85	rs1558744	intergenic	RPL39P28, IFNG	12	: 68,504,592	3.000 x 10-12	*	NHGRI	19122664
86	rs1558744	intergenic	RPL39P28, IFNG	12	: 68,504,592	4.000 x 10-12	*	NHGRI	20228799
87	rs3024505	intergenic	RPS14P1, IL10	1	: 206,939,904	6.000 x 10-17	**2b**	NHGRI	21297633
88	rs3024505	intergenic	RPS14P1, IL10	1	: 206,939,904	1.000 x 10-12	*	NHGRI	18836448
89	rs3024505	intergenic	RPS14P1, IL10	1	: 206,939,904	1.000 x 10-8	*	NHGRI	20228799
90	rs668853	intergenic	RPS2P34, RPS6P12	9	: 85,311,147	2.000 x 10-6	*	NHGRI	19122664
91	rs4510766	intergenic	SLC26A3, DLD	7	: 107,492,789	2.000 x 10-16	*	NHGRI	21297633
92	rs886774	intergenic	SLC26A3, DLD	7	: 107,495,434	3.000 x 10-8	*	NHGRI	19915572
93	rs4598195	intergenic	SLC26A3, DLD	7	: 107,503,441	8.000 x 10-8	*	NHGRI	20228799
94	rs2108225	intergenic	SLC26A3, DLD	7	: 107,453,103	1.000 x 10-7	*	NHGRI	19915573
95	rs4598195	intergenic	SLC26A3, DLD	7	: 107,503,441	1.000 x 10-7	*	NHGRI	19122664
96	rs4730273	intergenic	SLC26A3, DLD	7	: 107,479,519	5.000 x 10-7	*	NHGRI	19122664
97	rs4730276	intergenic	SLC26A3, DLD	7	: 107,484,437	9.000 x 10-7	*	NHGRI	19122664
98	rs4957048	intergenic	SLC9A3, CEP72	5	: 583,442	1.000 x 10-9	*	NHGRI	20228799
99	rs11739663	intergenic	SLC9A3, CEP72	5	: 594,083	3.000 x 10-8	*	NHGRI	21297633
100	rs7809799	intergenic	SMURF1, KPNA7	7	: 98,760,504	9.000 x 10-11	*	NHGRI	20228798
101	rs734999	intergenic	TNFRSF14, C1orf93	1	: 2,513,216	3.000 x 10-9	**1f**	NHGRI	21297633
102	rs943072	intergenic	VEGFA, C6orf223	6	: 43,795,968	2.000 x 10-10	*	NHGRI	21297633
103	rs7524102	intergenic	WNT4, ZBTB40	1	: 22,698,447	2.000 x 10-13	*	NHGRI	21297633
104	rs7524102	intergenic	WNT4, ZBTB40	1	: 22,698,447	3.000 x 10-7	*	NHGRI	19915572
105	rs6499188	intergenic	ZFP90, CDH3	16	: 68,674,788	4.000 x 10-8	*	NHGRI	21297633
106	rs2872507	intergenic	ZPBP2, GSDMB	17	: 38,040,763	5.000 x 10-11	*	NHGRI	21297633
107	rs8067378	intergenic	ZPBP2, GSDMB	17	: 38,051,348	1.000 x 10-7	**1f**	NHGRI	20228799

Psoriasis, PS (data only from NHGRI)

#	rs #	Context	Gene	Location	P-value	RegulomeDB score	Source	PubMed
1	rs2395029	UTR-3	HCP5	**6** : 31,431,780	2.000 x 10-26	*	NHGRI	18369459
2	rs33980500	missense	TRAF3IP2	**6** : 111,913,262	1.000 x 10-16	*	NHGRI	20953188
3	rs20541	missense	IL13	**5** : 131,995,964	5.000 x 10-15	*	NHGRI	19169254
4	rs12720356	missense	TYK2	**19** : 10,469,975	4.000 x 10-11	*	NHGRI	20953190
5	rs11209026	missense	IL23R	**1** : 67,705,958	7.000 x 10-7	*	NHGRI	20953190
6	rs3213094	intron	IL12B	**5** : 158,750,769	3.000 x 10-26	*	NHGRI	19169255
7	rs240993	intron	REV3L	**6** : 111,673,714	5.000 x 10-20	*	NHGRI	20953190
8	rs17716942	intron	KCNH7	**2** : 163,260,691	1.000 x 10-13	*	NHGRI	20953190
9	rs610604	intron	TNFAIP3	**6** : 138,199,417	9.000 x 10-12	*	NHGRI	19169254
10	rs27524	intron	CAST	**5** : 96,101,944	3.000 x 10-11	*	NHGRI	20953190
11	rs27524	intron	ERAP1	**5** : 96,101,944	3.000 x 10-11	*	NHGRI	20953190
12	rs4795067	intron	NOS2	**17** : 26,106,675	4.000 x 10-11	*	NHGRI	20953189
13	rs3213094	intron	IL12B	**5** : 158,750,769	5.000 x 10-11	*	NHGRI	20953190
14	rs10782001	intron	FBXL19	**16** : 30,942,625	9.000 x 10-10	*	NHGRI	20953189
15	rs2066808	intron	STAT2	**12** : 56,737,973	1.000 x 10-9	*	NHGRI	19169254
16	rs280519	intron	TYK2	**19** : 10,472,933	4.000 x 10-9	*	NHGRI	20953190
17	rs12586317	intron	KIAA0391	**14** : 35,682,172	2.000 x 10-8	*	NHGRI	20953189
18	rs2201841	intron	IL23R	**1** : 67,694,202	3.000 x 10-8	*	NHGRI	19169254
19	rs2066808	intron	STAT2	**12** : 56,737,973	2.000 x 10-7	*	NHGRI	20953190
20	rs610604	intron	TNFAIP3	**6** : 138,199,417	7.000 x 10-7	*	NHGRI	20953190
21	rs495337	cds-synon	SPATA2	**20** : 48,522,330	1.000 x 10-8	*	NHGRI	18364390
22	rs495337	cds-synon	SPATA2	**20** : 48,522,330	2.000 x 10-7	*	NHGRI	20953189
23	rs10484554	intergenic	WASF5P, HLA-B	**6** : 31,274,555	4.000 x 10-214	*	NHGRI	20953190
24	rs12191877	intergenic	RPL3P2, WASF5P	**6** : 31,252,925	1.000 x 10-100	*	NHGRI	19169254

25	rs10484554	intergenic	WASF5P, HLA-B	**6** : 31,274,555	2.000 x 10-39	*	NHGRI	18369459
26	rs12191877	intergenic	RPL3P2, WASF5P	**6** : 31,252,925	4.000 x 10-32	*	NHGRI	20953188
27	rs4085613	intergenic	LCE3E, LCE3D	**1** : 152,550,018	7.000 x 10-30	*	NHGRI	19169255
28	rs2082412	intergenic	UBLCP1, IL12B	**5** : 158,717,789	2.000 x 10-28	*	NHGRI	19169254
29	rs17728338	intergenic	TNIP1, ANXA6	**5** : 150,478,318	1.000 x 10-20	*	NHGRI	19169254
30	rs2546890	intergenic	IL12B, ADRA1B	**5** : 158,759,900	1.000 x 10-20	*	NHGRI	20953188
31	rs8016947	intergenic	PSMA6, RPLP0P3	**14** : 35,832,666	2.000 x 10-11	*	NHGRI	20953190
32	rs4112788	intergenic	LCE3E, LCE3D	**1** : 152,551,276	3.000 x 10-10	*	NHGRI	20953190
33	rs3134792	intergenic	WASF5P, HLA-B	**6** : 31,312,326	1.000 x 10-9	*	NHGRI	18364390
34	rs702873	intergenic	RPL21P33, REL	**2** : 61,081,542	4.000 x 10-9	*	NHGRI	20953190
35	rs4649203	intergenic	IL28RA, GRHL3	**1** : 24,519,920	7.000 x 10-8	*	NHGRI	20953190
36	rs6809854	intergenic	SATB1, KCNH8	**3** : 18,784,423	1.000 x 10-7	*	NHGRI	20953190
37	rs1975974	intergenic	C17orf51, UBBP4	**17** : 21,707,060	1.000 x 10-7	*	NHGRI	20953189
38	rs1008953	intergenic	SDC4, SYS1	**20** : 43,980,726	1.000 x 10-7	*	NHGRI	20953189
39	rs12580100	intergenic	RPS26, ERBB3	**12** : 56,439,209	1.000 x 10-6	**3a**	NHGRI	20953189

* all SNP are scored 4, 5 or 6 if not stated otherwise

AID associated GWAS SNPs retrieved from the NHGRI GWAS catalogue using PhenGenI software tool on the November 30th, 2014. Each SNP is identified by its dbSNP rs identifier along with its genomic context, associated genes with NCBI Gene identifier, location on a chromosome and PubMed ID.

Note: Regulome score is not retrieved from the Catalog; it is calculated as a part of our research additionally and will be discussed in the Section 3.3.3.

Supplemental Table 2. RegulomeDB scoring results for GWAS AID SNPs

Arthritis, Rheumatoid

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs4750316	UTR-3	DKFZp667F0711	2.000 x 10-6	
2	rs4750316	UTR-3	DKFZp667F0711	4.000 x 10-6	
3	rs1329568	UTR-3	LOC100130458	8.000 x 10-7	
4	rs805297	nearGene-5	APOM	3.000 x 10-10	**1f**
5	rs2812378	nearGene-5	CCL21	3.000 x 10-8	
6	rs2841277	nearGene-5	PLD4	2.000 x 10-14	
7	rs6496667	nearGene-5	ZNF774	1.000 x 10-6	
8	rs3087243	nearGene-3	CTLA4	1.000 x 10-8	**3a**
9	rs10488631	nearGene-3	TNPO3	4.000 x 10-11	**3a**
10	rs2233434	missense	NFKBIE	6.000 x 10-19	
11	rs2476601	missense	PTPN22	9.000 x 10-74	**2b**
12	rs2476601	missense	PTPN22	2.000 x 10-21	**2b**
13	rs2476601	missense	PTPN22	2.000 x 10-11	**2b**
14	rs3184504	missense	SH2B3	6.000 x 10-6	
15	rs2230926	missense	TNFAIP3	2.000 x 10-6	
16	rs2075876	intron	AIRE	4.000 x 10-9	
17	rs3816587	intron	ANAPC4	9.000 x 10-6	
18	rs6859219	intron	ANKRD55	1.000 x 10-11	**2b**
19	rs2867461	intron	ANXA3	1.000 x 10-12	
20	rs2062583	intron	ARHGEF3	2.000 x 10-6	
21	rs10821944	intron	ARID5B	6.000 x 10-18	
22	rs1600249	intron	BLK	5.000 x 10-6	
23	rs26232	intron	C5orf30	4.000 x 10-8	**2b**
24	rs6910071	intron	C6orf10	1.000 x 10-299	
25	rs2395148	intron	C6orf10	2.000 x 10-10	
26	rs3093024	intron	CCR6	8.000 x 10-19	6
27	rs3093023	intron	CCR6	2.000 x 10-11	**1f**
28	rs840016	intron	CD247	2.000 x 10-6	
29	rs4810485	intron	CD40	3.000 x 10-9	**1f**
30	rs4810485	intron	CD40	8.000 x 10-9	**1f**
31	rs42041	intron	CDK6	4.000 x 10-6	
32	rs4942242	intron	ENOX1	2.000 x 10-7	
33	rs1914816	intron	ETFA	7.000 x 10-7	
34	rs13315591	intron	FAM107A	5.000 x 10-8	
35	rs7940423	intron	GALNTL4	1.000 x 10-7	
36	rs3783637	intron	GCH1	2.000 x 10-6	**2b**
37	rs706778	intron	IL2RA	1.000 x 10-11	
38	rs13119723	intron	KIAA1109	7.000 x 10-7	
39	rs1678542	intron	KIF5A	9.000 x 10-8	

40	rs13393173	intron	LASS6	4.000 x 10-6	
41	rs17118552	intron	MDGA2	2.000 x 10-7	
42	rs3890745	intron	MMEL1	1.000 x 10-7	
43	rs3890745	intron	MMEL1	4.000 x 10-6	
44	rs7046653	intron	MOBKL2B	5.000 x 10-7	
45	rs6500395	intron	N4BP1	6.000 x 10-7	
46	rs3781913	intron	PDE2A	6.000 x 10-10	**1f**
47	rs2075876	intron	PFKL	4.000 x 10-9	
48	rs6026990	intron	PHACTR3	6.000 x 10-7	
49	rs854555	intron	PON1	2.000 x 10-6	
50	rs7404928	intron	PRKCB	4.000 x 10-6	
51	rs1957895	intron	PRKCH	4.000 x 10-7	
52	rs12901682	intron	PSMA4	4.000 x 10-8	
53	rs2847297	intron	PTPN2	2.000 x 10-8	
54	rs13137105	intron	RCHY1	9.000 x 10-7	
55	rs13031237	intron	REL	8.000 x 10-7	
56	rs16977065	intron	RIT2	1.000 x 10-7	
57	rs1809529	intron	SLC6A11	3.000 x 10-7	
58	rs934734	intron	SPRED2	5.000 x 10-10	
59	rs11121380	intron	SPSB1	5.000 x 10-8	
60	rs7574865	intron	STAT4	3.000 x 10-7	
61	rs7574865	intron	STAT4	2.000 x 10-6	
62	rs3761847	intron	TRAF1	4.000 x 10-14	
63	rs3761847	intron	TRAF1	2.000 x 10-7	
64	rs12831974	intron	TRHDE	6.000 x 10-6	**2b**
65	rs11203203	intron	UBASH3A	4.000 x 10-6	
66	rs12046117	intron	VTCN1	1.000 x 10-6	
67	rs1543922	intron	ZNF175	3.000 x 10-7	
68	rs2240335	cds-synon	PADI4	2.000 x 10-8	
69	rs2240335	cds-synon	PADI4	2.000 x 10-8	
70	rs11676922	intergenic	AFF3, LONRF2	1.000 x 10-14	
71	rs10865035	intergenic	AFF3, LONRF2	2.000 x 10-6	
72	rs1898036	intergenic	ATPBD4, COX6CP4	2.000 x 10-7	
73	rs11900673	intergenic	B3GNT2, TMEM17	1.000 x 10-8	
74	rs2002842	intergenic	BDP1P, SALL3	6.000 x 10-6	
75	rs11051970	intergenic	BICD1, FGD4	1.000 x 10-6	
76	rs12565755	intergenic	C1orf87, NFIA	5.000 x 10-8	
77	rs874040	intergenic	C4orf52, RBPJ	1.000 x 10-16	**2b**
78	rs951005	intergenic	C9orf144B, C9orf144	4.000 x 10-10	
79	rs11937061	intergenic	CCNG2, CXCL13	2.000 x 10-7	
80	rs657075	intergenic	CSF2, P4HA2	3.000 x 10-10	
81	rs6138150	intergenic	CST2, CST5	3.000 x 10-6	
82	rs1273516	intergenic	CYP4F22, RPL23AP2	9.000 x 10-7	
83	rs2837960	intergenic	DSCAM, C21orf130	2.000 x 10-6	
84	rs6138892	intergenic	EBF4, RPL19P1	3.000 x 10-7	

85	rs4937362	intergenic	ETS1, FLI1	8.000 x 10-7	
86	rs2736340	intergenic	FAM167A, BLK	6.000 x 10-9	
87	rs9604529	intergenic	FLJ44054	7.000 x 10-7	
88	rs16938910	intergenic	GDAP1, PCBP2P2	4.000 x 10-7	
89	rs12109285	intergenic	GUSBP1, CDH12	1.000 x 10-7	
90	rs1610677	intergenic	HCP5P12, HLA-G	4.000 x 10-15	
91	rs6457620	intergenic	HLA-DQB1, HLA-DQA2	4.000 x 10-186	
92	rs6457617	intergenic	HLA-DQB1, HLA-DQA2	5.000 x 10-75	
93	rs13192471	intergenic	HLA-DQB1, HLA-DQA2	2.000 x 10-58	
94	rs7765379	intergenic	HLA-DQB1, HLA-DQA2	5.000 x 10-23	
95	rs6457617	intergenic	HLA-DQB1, HLA-DQA2	1.000 x 10-9	
96	rs660895	intergenic	HLA-DRB1, HLA-DQA1	1.000 x 10-108	**1f**
97	rs9272219	intergenic	HLA-DRB1, HLA-DQA1	1.000 x 10-45	
98	rs615672	intergenic	HLA-DRB1, HLA-DQA1	8.000 x 10-27	
99	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	5.000 x 10-109	
100	rs6028945	intergenic	HSPE1P1, MAFB	2.000 x 10-7	
101	rs743777	intergenic	IL2RB, C1QTNF6	1.000 x 10-6	
102	rs743777	intergenic	IL2RB, C1QTNF6	2.000 x 10-6	
103	rs10488631	intergenic	IRF5, TNPO3	4.000 x 10-11	**3a**
104	rs2280381	intergenic	IRF8, FOXF1	2.000 x 10-6	
105	rs7155603	intergenic	JDP2, BATF	1.000 x 10-7	
106	rs231735	intergenic	KRT18P39, CTLA4	6.000 x 10-9	
107	rs983332	intergenic	LMO4, RPL36AP10	5.000 x 10-6	
108	rs6774280	intergenic	MRPS35P1, MRPS36P1	9.000 x 10-7	
109	rs9296015	intergenic	NOTCH4, C6orf10	2.000 x 10-38	
110	rs1406428	intergenic	NRXN1, CRYGGP	2.000 x 10-7	
111	rs6920220	intergenic	OLIG3, TNFAIP3	9.000 x 10-13	
112	rs10499194	intergenic	OLIG3, TNFAIP3	1.000 x 10-9	
113	rs6920220	intergenic	OLIG3, TNFAIP3	2.000 x 10-9	
114	rs6920220	intergenic	OLIG3, TNFAIP3	1.000 x 10-7	
115	rs1329568	intergenic	PAX5, RPL32P21	8.000 x 10-7	
116	rs16906916	intergenic	PCDH15, GAPDHP21	8.000 x 10-7	**3b**
117	rs881375	intergenic	PHF19, TRAF1	4.000 x 10-8	
118	rs12131057	intergenic	POU3F1, RRAGC	4.000 x 10-7	
119	rs10945919	intergenic	QKI, C6orf118	3.000 x 10-7	
120	rs13017599	intergenic	REL, RPS12P3	2.000 x 10-12	
121	rs12529514	intergenic	RNF182, CD83	2.000 x 10-8	
122	rs437943	intergenic	RPL31P31, ARAP2	4.000 x 10-6	
123	rs17374222	intergenic	RPLP1, GEMIN8P1	2.000 x 10-6	
124	rs11761231	intergenic	RPS14P10, RPS15AP22	4.000 x 10-7	
125	rs6679677	intergenic	RPS2P14, RSBN1	6.000 x 10-42	
126	rs6679677	intergenic	RPS2P14, RSBN1	6.000 x 10-25	
127	rs72991	intergenic	SC5DL, SORL1	5.000 x 10-7	
128	rs7164176	intergenic	SV2B, TRNAY16P	5.000 x 10-7	**3a**
129	rs800586	intergenic	TRPS1, EIF3H	2.000 x 10-7	

130	rs2872507	intergenic	ZPBP2, GSDMB	9.000 x 10-7	

Arthritis, Psoriatic

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs13191343	nearGene-5	HLA-C	2.000 x 10-72	
2	rs33980500	missense	TRAF3IP2	1.000 x 10-20	
3	rs12188300	intergenic	IL12B, ADRA1B	7.000 x 10-17	
4	rs13017599	intergenic	REL, RPS12P3	1.000 x 10-8	
5	rs702873	intergenic	RPL21P33, REL	2.000 x 10-7	

Spondylitis, Ankylosing

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs10781500	nearGene-5	CARD9	1.000 x 10-6	1f
2	rs30187	missense	ERAP1	2.000 x 10-27	
3	rs11209026	missense	IL23R	2.000 x 10-17	
4	rs11209026	missense	IL23R	9.000 x 10-14	
5	rs17095830	intron	ANO6	2.000 x 10-8	
6	rs4389526	intron	ANTXR2	9.000 x 10-8	
7	rs4333130	intron	ANTXR2	9.000 x 10-8	
8	rs1326986	intron	C10orf112	4.000 x 10-6	
9	rs2075726	intron	CSF2RB	9.000 x 10-6	
10	rs13210693	intron	FLJ37396	9.000 x 10-7	
11	rs2297909	intron	KIF21B	5.000 x 10-12	
12	rs27434	cds-synon	ERAP1	5.000 x 10-12	
13	rs10865331	intergenic	B3GNT2, TMEM17	7.000 x 10-34	
14	rs10865331	intergenic	B3GNT2, TMEM17	2.000 x 10-19	
15	rs10440635	intergenic	DAB2, PTGER4	3.000 x 10-7	
16	rs7743761	intergenic	DHFRP2, HLA-S	5.000 x 10-304	
17	rs13210693	intergenic	FLJ37396, CCDC162	9.000 x 10-7	2a
18	rs2242944	intergenic	FLJ45139, RPL23AP12	8.000 x 10-20	
19	rs378108	intergenic	FLJ45139, RPL23AP12	2.000 x 10-11	
20	rs4349859	intergenic	HLA-S, MICA	1.000 x 10-200	
21	rs6556416	intergenic	IL12B, ADRA1B	2.000 x 10-8	
22	rs2310173	intergenic	IL1R2, IL1R1	5.000 x 10-7	
23	rs8070463	intergenic	KPNB1, TBKBP1	5.000 x 10-8	
24	rs11616188	intergenic	LTBR, RPL31P10	4.000 x 10-12	
25	rs12146962	intergenic	MTCO1P2, NPAS3	9.000 x 10-6	
26	rs4552569	intergenic	RPL13AP14, EDIL3	9.000 x 10-10	
27	rs11249215	intergenic	RUNX3, SYF2	9.000 x 10-11	
28	rs1018326	intergenic	UBE2E3, ITGA4	2.000 x 10-6	

Psoriasis

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs8192583	UTR-5	GPSM3	2.250 x 10-20	1f
2	rs3094187	UTR-5	TCF19	3.725 x 10-13	1f
3	rs2240803	UTR-3	DPCR1	8.193 x 10-13	
4	rs2395029	UTR-3	HCP5	2.000 x 10-26	
5	rs8365	UTR-3	RNF5	1.094 x 10-17	1f
6	rs3130453	STOP-GAIN	CCHCR1	2.057 x 10-12	
7	rs3132965	nearGene-5	AGPAT1	8.389 x 10-16	
8	rs3094187	nearGene-5	CCHCR1	3.725 x 10-13	
9	rs7773175	nearGene-5	HLA-C	4.772 x 10-30	
10	rs2249742	nearGene-5	HLA-C	6.346 x 10-18	1f
11	rs2395471	nearGene-5	HLA-C	8.836 x 10-18	1f
12	rs2249741	nearGene-5	HLA-C	3.340 x 10-12	
13	rs2524082	nearGene-5	HLA-C	2.215 x 10-11	
14	rs9267502	nearGene-5	LST1	3.786 x 10-19	
15	rs2734573	nearGene-5	MCCD1	1.499 x 10-12	
16	rs176095	nearGene-5	PBX2	2.183 x 10-12	1f
17	rs3130453	nearGene-5	TCF19	2.057 x 10-12	
18	rs2021723	nearGene-5	TRIM40	8.690 x 10-15	
19	rs8365	nearGene-3	AGER	1.094 x 10-17	
20	rs1265086	nearGene-3	CCHCR1	2.689 x 10-11	1f
21	rs9468843	nearGene-3	DDR1	9.009 x 10-14	
22	rs176095	nearGene-3	GPSM3	2.183 x 10-12	
23	rs2853950	nearGene-3	HLA-C	1.756 x 10-14	
24	rs11575907	missense	HLA-DOB	5.960 x 10-19	
25	rs20541	missense	IL13	5.000 x 10-15	
26	rs11209026	missense	IL23R	7.000 x 10-7	
27	rs8192591	missense	NOTCH4	3.403 x 10-24	
28	rs33980500	missense	TRAF3IP2	1.000 x 10-16	
29	rs12720356	missense	TYK2	4.000 x 10-11	
30	rs2295663	intron	BAT5	8.396 x 10-24	
31	rs9267673	intron	C2	2.300 x 10-33	1f
32	rs6906662	intron	C6orf10	2.109 x 10-14	
33	rs27524	intron	CAST	3.000 x 10-11	
34	rs1265078	intron	CCHCR1	2.039 x 10-21	
36	rs2239518	intron	DDR1	3.454 x 10-12	2b
37	rs27524	intron	ERAP1	3.000 x 10-11	
38	rs10782001	intron	FBXL19	9.000 x 10-10	
39	rs3213094	intron	IL12B	3.000 x 10-26	
40	rs3213094	intron	IL12B	5.000 x 10-11	
41	rs2201841	intron	IL23R	3.000 x 10-8	
42	rs17716942	intron	KCNH7	1.000 x 10-13	
43	rs12586317	intron	KIAA0391	2.000 x 10-8	

44	rs9295938	intron	MUC21	9.163 x 10-18	
45	rs4795067	intron	NOS2	4.000 x 10-11	
46	rs3823418	intron	PSORS1C1	1.169 x 10-33	
47	rs3094205	intron	PSORS1C1	1.733 x 10-15	
48	rs3130573	intron	PSORS1C2	6.212 x 10-15	**1f**
49	rs3130573	intron	PSORS1C2	6.212 x 10-15	**1f**
50	rs240993	intron	REV3L	5.000 x 10-20	
51	rs3132965	intron	RNF5	8.389 x 10-16	
52	rs2066808	intron	STAT2	1.000 x 10-9	
53	rs2066808	intron	STAT2	2.000 x 10-7	
54	rs3093662	intron	TNF	4.752 x 10-21	
55	rs610604	intron	TNFAIP3	9.000 x 10-12	
56	rs610604	intron	TNFAIP3	7.000 x 10-7	
57	rs2077580	intron	TNXB	7.445 x 10-32	
58	rs2107195	intron	TRIM15	1.675 x 10-17	
60	rs280519	intron	TYK2	4.000 x 10-9	
61	rs8192583	cds-synon	NOTCH4	2.250 x 10-20	
62	rs443198	intron	NOTCH4	1.526 x 10-11	**1f**
63	rs495337	cds-synon	SPATA2	1.000 x 10-8	
64	rs495337	cds-synon	SPATA2	2.000 x 10-7	
65	rs7762370	intergenic	BTNL2, HLA-DRA	5.130 x 10-19	
66	rs9501624	intergenic	BTNL2, HLA-DRA	2.357 x 10-12	
67	rs1975974	intergenic	C17orf51, UBBP4	1.000 x 10-7	
68	rs3130955	intergenic	HCG22, C6orf15	1.312 x 10-20	**1f**
69	rs2844627	intergenic	HCG27, HLA-C	2.307 x 10-36	
70	rs2394895	intergenic	HCG27, HLA-C	2.593 x 10-29	
71	rs3130517	intergenic	HCG27, HLA-C	1.430 x 10-28	
72	rs3130713	intergenic	HCG27, HLA-C	9.808 x 10-28	
73	rs3130467	intergenic	HCG27, HLA-C	1.339 x 10-27	
74	rs9263967	intergenic	HCG27, HLA-C	7.296 x 10-15	
75	rs3130685	intergenic	HCG27, HLA-C	8.583 x 10-14	
76	rs3130425	intergenic	HCG27, HLA-C	1.808 x 10-12	
77	rs3095250	intergenic	HCG27, HLA-C	2.892 x 10-12	
78	rs3132496	intergenic	HCG27, HLA-C	2.898 x 10-12	
79	rs3132486	intergenic	HLA-C, RPL3P2	6.660 x 10-13	
80	rs3132485	intergenic	HLA-C, RPL3P2	2.623 x 10-12	
81	rs2647087	intergenic	HLA-DQB1, HLA-DQA2	1.865 x 10-16	
82	rs2858333	intergenic	HLA-DQB1, HLA-DQA2	8.551 x 10-16	
83	rs2856726	intergenic	HLA-DQB1, HLA-DQA2	5.086 x 10-12	
84	rs12203586	intergenic	HLA-DQB1, HLA-DQA2	5.743 x 10-12	**1f**
85	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	8.870 x 10-13	
86	rs10484552	intergenic	HLA-E, GNL1	7.487 x 10-19	
87	rs13437088	intergenic	HLA-S, MICA	2.539 x 10-13	
88	rs10947208	intergenic	HLA-S, MICA	2.182 x 10-12	
89	rs7756521	intergenic	IER3, DDR1	7.581 x 10-19	

90	rs4711229	intergenic	IER3, DDR1	5.431 x 10-14	
91	rs9295917	intergenic	IER3, DDR1	5.091 x 10-13	
92	rs4713380	intergenic	IER3, DDR1	6.508 x 10-13	**2b**
93	rs12526186	intergenic	IER3, DDR1	1.272 x 10-12	
94	rs13198118	intergenic	IER3, DDR1	2.074 x 10-12	**2b**
95	rs7749924	intergenic	IER3, DDR1	2.249 x 10-12	
96	rs9295924	intergenic	IER3, DDR1	2.685 x 10-12	
97	rs3131043	intergenic	IER3, DDR1	1.814 x 10-11	**1f**
98	rs2546890	intergenic	IL12B, ADRA1B	1.000 x 10-20	
99	rs4649203	intergenic	IL28RA, GRHL3	7.000 x 10-8	
100	rs4085613	intergenic	LCE3E, LCE3D	7.000 x 10-30	
101	rs4112788	intergenic	LCE3E, LCE3D	3.000 x 10-10	
102	rs9501106	intergenic	MICA, HLA-X	3.527 x 10-16	
103	rs9295993	intergenic	MICA, HLA-X	1.344 x 10-15	
104	rs9266844	intergenic	MICA, HLA-X	1.022 x 10-11	
105	rs7772549	intergenic	MICA, HLA-X	1.032 x 10-11	
106	rs9266846	intergenic	MICA, HLA-X	2.441 x 10-11	
107	rs17476793	intergenic	MICC, SUCLA2P1	1.261 x 10-17	
108	rs13191258	intergenic	MUC21, HCG22	3.951 x 10-33	
109	rs2844645	intergenic	MUC21, HCG22	7.292 x 10-19	
110	rs9366764	intergenic	MUC21, HCG22	8.095 x 10-18	
111	rs9394031	intergenic	MUC21, HCG22	8.254 x 10-18	
112	rs3871466	intergenic	MUC21, HCG22	7.904 x 10-17	
113	rs2523870	intergenic	MUC21, HCG22	2.725 x 10-12	
114	rs2517552	intergenic	MUC21, HCG22	8.159 x 10-12	
115	rs2894176	intergenic	MUC21, HCG22	9.713 x 10-12	
116	rs2517527	intergenic	MUC21, HCG22	1.831 x 10-11	
117	rs9262492	intergenic	MUC21, HCG22	2.993 x 10-11	**2b**
118	rs6916062	intergenic	NOTCH4, C6orf10	2.524 x 10-22	
119	rs9267463	intergenic	PPIAP9, RPL15P4	1.104 x 10-23	
120	rs9267464	intergenic	PPIAP9, RPL15P4	1.806 x 10-22	
121	rs2734573	intergenic	PPIAP9, RPL15P4	1.499 x 10-12	
122	rs8016947	intergenic	PSMA6, RPLP0P3	2.000 x 10-11	
123	rs702873	intergenic	RPL21P33, REL	4.000 x 10-9	
124	rs842636	intergenic	RPL21P33, REL	6.000 x 10-6	
125	rs12191877	intergenic	RPL3P2, WASF5P	1.000 x 10-100	
126	rs12191877	intergenic	RPL3P2, WASF5P	2.024 x 10-51	
127	rs12191877	intergenic	RPL3P2, WASF5P	4.000 x 10-32	
128	rs12580100	intergenic	RPS26, ERBB3	1.000 x 10-7	**3a**
129	rs6809854	intergenic	SATB1, KCNH8	1.000 x 10-7	
130	rs1008953	intergenic	SDC4, SYS1	1.000 x 10-7	
131	rs17728338	intergenic	TNIP1, ANXA6	1.000 x 10-20	
132	rs1015465	intergenic	TRIM31, TRIM40	3.685 x 10-15	
133	rs2082412	intergenic	UBLCP1, IL12B	2.000 x 10-28	
134	rs10484554	intergenic	WASF5P, HLA-B	4.000 x 10-214	

135	rs9468933	intergenic	WASF5P, HLA-B	1.621 x 10-46	
136	rs10484554	intergenic	WASF5P, HLA-B	2.000 x 10-39	
137	rs2894207	intergenic	WASF5P, HLA-B	3.766 x 10-38	
138	rs9380237	intergenic	WASF5P, HLA-B	2.576 x 10-28	
139	rs2524163	intergenic	WASF5P, HLA-B	5.329 x 10-20	
140	rs2243868	intergenic	WASF5P, HLA-B	2.798 x 10-19	**1f**
141	rs2853923	intergenic	WASF5P, HLA-B	2.701 x 10-15	
142	rs9380240	intergenic	WASF5P, HLA-B	4.412 x 10-14	
143	rs2442719	intergenic	WASF5P, HLA-B	2.809 x 10-13	
144	rs3873386	intergenic	WASF5P, HLA-B	1.173 x 10-12	**1f**
145	rs2156875	intergenic	WASF5P, HLA-B	3.707 x 10-12	**1f**
146	rs9468937	intergenic	WASF5P, HLA-B	4.828 x 10-12	
147	rs3134792	intergenic	WASF5P, HLA-B	1.000 x 10-9	
148	rs9267673	intergenic	ZBTB12, C2	2.300 x 10-33	

Crohn Disease

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs504963	UTR-3	FUT2	2.000 x 10-8	
2	rs10210302	nearGene-5	ATG16L1	5.000 x 10-14	
3	rs12677663	nearGene-5	C8orf84	2.000 x 10-8	
4	rs11190140	nearGene-5	NKX2-3	3.000 x 10-16	
5	rs11190141	nearGene-5	NKX2-3	5.000 x 10-7	
6	rs11574514	nearGene-5	PSMB10	2.000 x 10-7	
7	rs2241880	missense	ATG16L1	1.000 x 10-13	
8	rs2241880	missense	ATG16L1	1.000 x 10-12	
9	rs2241880	missense	ATG16L1	3.000 x 10-6	
10	rs3764147	missense	C13orf31 (LACC1)	2.000 x 10-13	**1f**
11	rs4077515	missense	CARD9	1.000 x 10-36	**1f**
12	rs11209026	missense	IL23R	1.000 x 10-64	
13	rs11209026	missense	IL23R	4.000 x 10-21	
14	rs11209026	missense	IL23R	1.000 x 10-18	
15	rs11209026	missense	IL23R	2.000 x 10-18	
16	rs3197999	missense	MST1	6.000 x 10-17	
17	rs3197999	missense	MST1	1.000 x 10-12	
18	rs2476601	missense	PTPN22	1.000 x 10-8	**2b**
19	rs12720356	missense	TYK2	1.000 x 10-12	
20	rs3792109	intron	ATG16L1	7.000 x 10-41	
21	rs3828309	intron	ATG16L1	2.000 x 10-32	
22	rs1847472	intron	BACH2	5.000 x 10-9	
23	rs102275	intron	C11orf10	2.000 x 10-11	
24	rs12521868	intron	C5orf56	1.000 x 10-20	
25	rs2188962	intron	C5orf56	2.000 x 10-18	**2a**
26	rs2188962	intron	C5orf56	1.000 x 10-7	
27	rs6908425	intron	CDKAL1	9.000 x 10-10	

28	rs151181	intron	CLN3	2.000 x 10-11	
29	rs1998598	intron	DENND1B	9.000 x 10-9	
30	rs13428812	intron	DNMT3A	9.000 x 10-10	
31	rs2549794	intron	ERAP2	1.000 x 10-10	**1f**
32	rs2301436	intron	FGFR1OP	1.000 x 10-12	
33	rs2301436	intron	FGFR1OP	6.000 x 10-8	
34	rs780093	intron	GCKR	5.000 x 10-11	
35	rs8005161	intron	GPR65	4.000 x 10-18	
36	rs2058660	intron	IL18RAP	2.000 x 10-12	
37	rs11465804	intron	IL23R	7.000 x 10-63	
38	rs7517847	intron	IL23R	3.000 x 10-12	
39	rs11805303	intron	IL23R	6.000 x 10-12	
40	rs11465804	intron	IL23R	1.000 x 10-6	
41	rs12722489	intron	IL2RA	3.000 x 10-9	
42	rs2274910	intron	ITLN1	1.000 x 10-9	
43	rs1793004	intron	NELL1	3.000 x 10-6	
44	rs2076756	intron	NOD2	4.000 x 10-69	
45	rs2076756	intron	NOD2	1.000 x 10-37	
46	rs2076756	intron	NOD2	1.000 x 10-21	
47	rs5743289	intron	NOD2	6.000 x 10-17	
48	rs2076756	intron	NOD2	7.000 x 10-14	
49	rs17221417	intron	NOD2	4.000 x 10-11	
50	rs5743289	intron	NOD2	1.000 x 10-6	
51	rs2797685	intron	PER3	7.000 x 10-9	
52	rs6738825	intron	PLCL1	4.000 x 10-9	
53	rs13003464	intron	PUS10	5.000 x 10-9	
54	rs10181042	intron	PUS10	7.000 x 10-9	
55	rs17309827	intron	SLC22A23	7.000 x 10-9	
56	rs17293632	intron	SMAD3	3.000 x 10-19	**2a**
57	rs7423615	intron	SP140	3.000 x 10-13	
58	rs744166	intron	STAT3	7.000 x 10-12	
59	rs10495903	intron	THADA	2.000 x 10-14	
60	rs4263839	intron	TNFSF15	3.000 x 10-10	**2b**
61	rs181359	intron	UBE2L3	5.000 x 10-16	**1f**
62	rs4809330	intron	ZGPAT	3.000 x 10-15	
63	rs1250550	intron	ZMIZ1	1.000 x 10-30	
64	rs7076156	intron	ZNF365	7.000 x 10-9	
65	rs2066847	frameshift	NOD2	3.000 x 10-24	
66	rs2066847	frameshift	NOD2	2.000 x 10-15	
67	rs9858542	cds-synon	BSN	4.000 x 10-8	**1f**
68	rs9858542	cds-synon	BSN	5.000 x 10-8	
69	rs1142287	cds-synon	SCAMP3	2.000 x 10-13	
70	rs3810936	cds-synon	TNFSF15	1.000 x 10-15	
71	rs6545946	intergenic	B3GNT2, TMEM17	7.000 x 10-9	
72	rs359457	intergenic	BOD1, CPEB4	3.000 x 10-12	

73	rs1398024	intergenic	C10orf67, OTUD1	4.000 x 10-6	
74	rs7927894	intergenic	C11orf30, LRRC32	1.000 x 10-9	
75	rs11584383	intergenic	C1orf81, KIF21B	1.000 x 10-11	
76	rs762421	intergenic	C21orf33, ICOSLG	1.000 x 10-9	
77	rs13361189	intergenic	C5orf62, IRGM	2.000 x 10-10	
78	rs1456893	intergenic	C7orf72, IKZF1	5.000 x 10-9	
79	rs3091315	intergenic	CCL2, CCL7	2.000 x 10-13	
80	rs3091316	intergenic	CCL2, CCL7	4.000 x 10-8	
81	rs7807268	intergenic	CNTNAP2, RPL32P17	4.000 x 10-6	
82	rs11742570	intergenic	DAB2, PTGER4	7.000 x 10-36	
83	rs4613763	intergenic	DAB2, PTGER4	7.000 x 10-27	
84	rs9292777	intergenic	DAB2, PTGER4	3.000 x 10-18	
85	rs17234657	intergenic	DAB2, PTGER4	2.000 x 10-12	
86	rs1373692	intergenic	DAB2, PTGER4	2.000 x 10-12	
87	rs9292777	intergenic	DAB2, PTGER4	2.000 x 10-11	
88	rs1992660	intergenic	DAB2, PTGER4	4.000 x 10-7	
89	rs2062305	intergenic	FABP3P2, TNFSF11	5.000 x 10-10	
90	rs10801047	intergenic	FAM5C, RGS18	3.000 x 10-8	
91	rs9286879	intergenic	FASLG, TNFSF18	2.000 x 10-9	
92	rs12035082	intergenic	FASLG, TNFSF18	2.000 x 10-7	
93	rs2836754	intergenic	FLJ45139, RPL23AP12	5.000 x 10-7	
94	rs281379	intergenic	FUT2, MAMSTR	7.000 x 10-12	
95	rs4409764	intergenic	GOT1, NKX2-3	2.000 x 10-20	**3a**
96	rs10883365	intergenic	GOT1, NKX2-3	4.000 x 10-10	
97	rs10883365	intergenic	GOT1, NKX2-3	6.000 x 10-8	
98	rs9469220	intergenic	HLA-DQB1, HLA-DQA2	2.000 x 10-6	**1f**
99	rs9258260	intergenic	IFITM4P, 3.8-1.5	2.000 x 10-10	**1f**
100	rs10045431	intergenic	IL12B, ADRA1B	4.000 x 10-13	
101	rs10045431	intergenic	IL12B, ADRA1B	7.000 x 10-8	
102	rs6887695	intergenic	IL12B, ADRA1B	9.000 x 10-6	
103	rs3091338	intergenic	IL3, CSF2	4.000 x 10-8	
104	rs7714584	intergenic	IRGM, ZNF300	8.000 x 10-19	
105	rs11747270	intergenic	IRGM, ZNF300	3.000 x 10-16	
106	rs1000113	intergenic	IRGM, ZNF300	3.000 x 10-7	
107	rs6601764	intergenic	KLF6, AKR1E2	9.000 x 10-6	
108	rs11167764	intergenic	MRPL11P2, NDFIP1	2.000 x 10-9	
109	rs1819658	intergenic	MRPS35P3, IPMK	9.000 x 10-17	
110	rs1736135	intergenic	NRIP1, CYCSP42	7.000 x 10-9	
111	rs2413583	intergenic	PDGFB, RPL3	1.000 x 10-26	
112	rs694739	intergenic	PRDX5, CCDC88B	6.000 x 10-10	**1f**
113	rs2542151	intergenic	PSMG2, PTPN2	5.000 x 10-17	
114	rs2542151	intergenic	PSMG2, PTPN2	3.000 x 10-8	
115	rs2542151	intergenic	PSMG2, PTPN2	2.000 x 10-7	
116	rs6651252	intergenic	PVT1, GSDMC	4.000 x 10-18	
117	rs10758669	intergenic	RCL1, JAK2	3.000 x 10-9	

118	rs4902642	intergenic	RPL12P7, ZFP36L1	2.000 x 10-10	
119	rs11175593	intergenic	RPL30P13, LRRK2	3.000 x 10-10	
120	rs7746082	intergenic	RPL35P3, PRDM1	2.000 x 10-10	
121	rs17582416	intergenic	RPS12P16, CUL2	2.000 x 10-9	
122	rs3024505	intergenic	RPS14P1, IL10	2.000 x 10-14	**2b**
123	rs713875	intergenic	RPS3AP51, LIF	7.000 x 10-12	**1f**
124	rs13073817	intergenic	SATB1, KCNH8	7.000 x 10-9	
125	rs6596075	intergenic	SLC22A5, C5orf56	3.000 x 10-6	**1f**
126	rs11229030	intergenic	SLC43A3, RTN4RL2	8.000 x 10-9	
127	rs736289	intergenic	SLC7A10, CEBPA	9.000 x 10-9	
128	rs7705924	intergenic	SLCO6A1, PAM	2.000 x 10-8	
129	rs212388	intergenic	TAGAP, FNDC1	2.000 x 10-11	
130	rs10734105	intergenic	TCERG1L, FLJ46300	3.000 x 10-8	
131	rs7702331	intergenic	TMEM174, FOXD1	6.000 x 10-12	
132	rs1551398	intergenic	TRIB1, FAM84B	5.000 x 10-9	
133	rs1906493	intergenic	TRIB1, FAM84B	3.000 x 10-6	
134	rs10761659	intergenic	ZNF365, ALDH7A1P4	4.000 x 10-22	
135	rs10995271	intergenic	ZNF365, ALDH7A1P4	4.000 x 10-20	
136	rs224136	intergenic	ZNF365, ALDH7A1P4	1.000 x 10-10	
137	rs10761659	intergenic	ZNF365, ALDH7A1P4	2.000 x 10-6	
138	rs2872507	intergenic	ZPBP2, GSDMB	5.000 x 10-9	

Colitis, Ulcerative

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs10889677	UTR-3	IL23R	1.000 x 10-8	
2	rs2297441	UTR-3	RTEL1	2.000 x 10-10	
3	rs10781500	nearGene-5	CARD9	7.000 x 10-6	**1f**
4	rs678170	nearGene-5	FAM55A	5.000 x 10-14	
5	rs907611	nearGene-5	LSP1	1.000 x 10-10	**2a**
6	rs11190140	nearGene-5	NKX2-3	1.000 x 10-8	
7	rs3806308	nearGene-5	RNF186	7.000 x 10-9	**2b**
8	rs2297441	nearGene-5	TNFRSF6B (it is acually RTEL1)	2.000 x 10-10	
9	rs1317209	nearGene-3	RNF186	2.000 x 10-10	
10	rs4077515	missense	CARD9	5.000 x 10-8	**1f**
11	rs1801274	missense	FCGR2A	2.000 x 10-20	
12	rs1801274	missense	FCGR2A	2.000 x 10-12	
13	rs2305480	missense	GSDMB	3.000 x 10-8	
14	rs5771069	missense	IL17REL	4.000 x 10-8	
15	rs5771069	missense	IL17REL	2.000 x 10-7	
16	rs11209026	missense	IL23R	5.000 x 10-28	
17	rs11209026	missense	IL23R	3.000 x 10-10	
18	rs11209026	missense	IL23R	1.000 x 10-8	

19	rs3194051	missense	IL7R	4.000 x 10-8	
20	rs3197999	missense	MST1	4.000 x 10-9	
21	rs17388568	intron	ADAD1	9.000 x 10-7	
22	rs9822268	intron	APEH	2.000 x 10-17	
23	rs7554511	intron	C1orf106	2.000 x 10-13	
24	rs7554511	intron	C1orf106	1.000 x 10-6	
25	rs9263739	intron	CCHCR1	4.000 x 10-67	**1f**
26	rs12261843	intron	CCNY	7.000 x 10-10	
27	rs4781011	intron	CIITA	3.000 x 10-6	**2b**
28	rs267939	intron	DAP	6.000 x 10-12	
29	rs798502	intron	GNA12	3.000 x 10-15	**1b**
30	rs3024493	intron	IL10	1.000 x 10-12	**2b**
31	rs3024493	intron	IL10	8.000 x 10-8	**2b**
32	rs2201841	intron	IL23R	1.000 x 10-13	
33	rs2870946	intron	IL26	5.000 x 10-7	
34	rs2158836	intron	LAMB1	7.000 x 10-6	
35	rs4654925	intron	OTUD3	9.000 x 10-22	
36	rs35675666	intron	PARK7	5.000 x 10-9	
37	rs7608910	intron	PUS10	2.000 x 10-14	
38	rs13003464	intron	PUS10	7.000 x 10-9	
39	rs1992950	intron	SATB2	5.000 x 10-6	
40	rs4246905	intron	TNFSF15	6.000 x 10-12	
41	rs1728785	intron	ZFP90	3.000 x 10-8	
42	rs9858542	cds-synon	BSN	7.000 x 10-9	**1f**
43	rs9268480	cds-synon	BTNL2	3.000 x 10-6	
44	rs10781499	cds-synon	CARD9	3.000 x 10-19	**1f**
45	rs2155219	intergenic	C11orf30, LRRC32	5.000 x 10-16	
46	rs10800309	intergenic	C1orf192, FCGR2A	3.000 x 10-9	
47	rs11584383	intergenic	C1orf81, KIF21B	2.000 x 10-7	
48	rs2838519	intergenic	C21orf33, ICOSLG	6.000 x 10-11	
49	rs941823	intergenic	COG6, FOXO1	4.000 x 10-12	
50	rs9548988	intergenic	COG6, FOXO1	3.000 x 10-7	
51	rs11676348	intergenic	CXCR2, CXCR1	1.000 x 10-10	**1f**
52	rs6451493	intergenic	DAB2, PTGER4	3.000 x 10-9	**3a**
53	rs2836878	intergenic	FLJ45139, RPL23AP12	2.000 x 10-22	
54	rs6584283	intergenic	GOT1, NKX2-3	8.000 x 10-21	
55	rs6584283	intergenic	GOT1, NKX2-3	2.000 x 10-7	
56	rs6584283	intergenic	GOT1, NKX2-3	2.000 x 10-6	
57	rs17085007	intergenic	GPR12, RPS20P32	1.000 x 10-16	
58	rs17085007	intergenic	GPR12, RPS20P32	7.000 x 10-8	
59	rs4676406	intergenic	GPR35, AQP12B	8.000 x 10-11	
60	rs9268853	intergenic	HLA-DRB9, HLA-DRB5	1.000 x 10-55	
61	rs9268877	intergenic	HLA-DRB9, HLA-DRB5	4.000 x 10-23	

62	rs2395185	intergenic	HLA-DRB9, HLA-DRB5	5.000 x 10-22	
63	rs9268877	intergenic	HLA-DRB9, HLA-DRB5	6.000 x 10-18	
64	rs2395185	intergenic	HLA-DRB9, HLA-DRB5	1.000 x 10-16	
65	rs9268923	intergenic	HLA-DRB9, HLA-DRB5	4.000 x 10-15	
66	rs6017342	intergenic	HNF4A, RPL37AP1	1.000 x 10-20	
67	rs6017342	intergenic	HNF4A, RPL37AP1	9.000 x 10-17	
68	rs6871626	intergenic	IL12B, ADRA1B	1.000 x 10-21	
69	rs2310173	intergenic	IL1R2, IL1R1	3.000 x 10-12	
70	rs10975003	intergenic	INSL6, INSL4	1.000 x 10-6	
71	rs16940202	intergenic	IRF8, FOXF1	6.000 x 10-19	
72	rs4728142	intergenic	KCP, IRF5	2.000 x 10-8	**1f**
73	rs1297265	intergenic	NRIP1, CYCSP42	7.000 x 10-13	**2a**
74	rs1736135	intergenic	NRIP1, CYCSP42	2.000 x 10-7	
75	rs6920220	intergenic	OLIG3, TNFAIP3	8.000 x 10-17	
76	rs254560	intergenic	PITX1, H2AFY	1.000 x 10-9	
77	rs10758669	intergenic	RCL1, JAK2	2.000 x 10-25	
78	rs10758669	intergenic	RCL1, JAK2	1.000 x 10-6	
79	rs6426833	intergenic	RNF186, OTUD3	4.000 x 10-35	
80	rs6426833	intergenic	RNF186, OTUD3	2.000 x 10-21	
81	rs6426833	intergenic	RNF186, OTUD3	5.000 x 10-13	
82	rs6426833	intergenic	RNF186, OTUD3	2.000 x 10-11	
83	rs6911490	intergenic	RPL35P3, PRDM1	1.000 x 10-8	
84	rs7134599	intergenic	RPL39P28, IFNG	1.000 x 10-16	
85	rs1558744	intergenic	RPL39P28, IFNG	3.000 x 10-12	
86	rs1558744	intergenic	RPL39P28, IFNG	4.000 x 10-12	
87	rs3024505	intergenic	RPS14P1, IL10	6.000 x 10-17	**2b**
88	rs3024505	intergenic	RPS14P1, IL10	1.000 x 10-12	
89	rs3024505	intergenic	RPS14P1, IL10	1.000 x 10-8	
90	rs668853	intergenic	RPS2P34, RPS6P12	2.000 x 10-6	
91	rs4510766	intergenic	SLC26A3, DLD	2.000 x 10-16	
92	rs886774	intergenic	SLC26A3, DLD	3.000 x 10-8	
93	rs4598195	intergenic	SLC26A3, DLD	8.000 x 10-8	
94	rs2108225	intergenic	SLC26A3, DLD	1.000 x 10-7	
95	rs4598195	intergenic	SLC26A3, DLD	1.000 x 10-6	
96	rs4730273	intergenic	SLC26A3, DLD	5.000 x 10-6	
97	rs4730276	intergenic	SLC26A3, DLD	9.000 x 10-6	
98	rs4957048	intergenic	SLC9A3, CEP72	1.000 x 10-9	
99	rs11739663	intergenic	SLC9A3, CEP72	3.000 x 10-8	
100	rs7809799	intergenic	SMURF1, KPNA7	9.000 x 10-11	
101	rs734999	intergenic	TNFRSF14, C1orf93	3.000 x 10-9	**1f**
102	rs943072	intergenic	VEGFA, C6orf223	2.000 x 10-10	
103	rs7524102	intergenic	WNT4, ZBTB40	2.000 x 10-13	
104	rs7524102	intergenic	WNT4, ZBTB40	3.000 x 10-7	

105	rs6499188	intergenic	ZFP90, CDH3	4.000 x 10-8	
106	rs2872507	intergenic	ZPBP2, GSDMB	5.000 x 10-11	
107	rs8067378	intergenic	ZPBP2, GSDMB	1.000 x 10-7	**1f**

Inflammatory Bowel Diseases

#	rs #	Context	Gene	P-value	RegulomeDB score
1	rs10889677	UTR-3	IL23R	9.037 x 10-11	
2	rs11209026	missense	IL23R	4.000 x 10-11	
3	rs11209026	missense	IL23R	4.592 x 10-11	
4	rs11209026	missense	IL23R	7.000 x 10-11	
5	rs2315008	intron	ZGPAT	9.000 x 10-15	
6	rs2076756	intron	NOD2	1.262 x 10-14	**5**
7	rs7517847	intron	IL23R	2.991 x 10-13	
8	rs7517847	intron	IL23R	4.000 x 10-13	
9	rs1343151	intron	IL23R	1.628 x 10-11	
10	rs10489629	intron	IL23R	6.790 x 10-11	
11	rs2201841	intron	IL23R	3.574 x 10-10	
12	rs11465804	intron	IL23R	3.737 x 10-10	
13	rs5743289	intron	NOD2	4.000 x 10-10	
14	rs2076756	intron	NOD2	5.000 x 10-10	
15	rs1004819	intron	IL23R	1.504 x 10-9	
16	rs8049439	intron	ATXN2L	2.000 x 10-9	**1b**
17	rs2412973	intron	HORMAD2	2.000 x 10-9	
18	rs1250550	intron	ZMIZ1	6.000 x 10-9	
19	rs2066843	cds-synon	NOD2	7.869 x 10-13	
20	rs9271366	intergenic	HLA-DRB1, HLA-DQA1	2.000 x 10-70	
21	rs9271366	intergenic	HLA-DRB1, HLA-DQA1	3.000 x 10-31	
22	rs2006996	intergenic	TNFSF15, TNFSF8	4.000 x 10-16	
23	rs2006996	intergenic	TNFSF15, TNFSF8	4.000 x 10-13	
24	rs2836878	intergenic	FLJ45139, RPL23AP12	4.000 x 10-12	
25	rs9271366	intergenic	HLA-DRB1, HLA-DQA1	8.000 x 10-11	
26	rs10500264	intergenic	SLC7A10, CEBPA	4.000 x 10-10	
27	rs11209032	intergenic	IL23R, IL12RB2	8.645 x 10-10	
28	rs477515	intergenic	HLA-DRB1, HLA-DQA1	1.000 x 10-8	

Note:

prvided are only the highest scores, and two additional interesting SNPs; otherwise all SNPs without numbers have scores higher than 3: 4, 5, 6 or no data existed at the time of data retreival

AID associated GWAS SNPs, labeled with its rs identifier, are associated with genes; for each SNP its RegulomeDB score is given. Only highest scored SNPs are further analysed. ncSNPs are sorted by the associated AID.

Supplemetal Table 3.

Non-coding AID GWAS SNPs might influence the following genes and their corresponding KEGG pathways:

	Gene	GeneID	KEGG pathway		Gene function, association or other	Directly involved in Immune System (and not present ubiquitously)
1	ANKRD55	79722	none		ankyrin repeat domain 55;linked to MS, insulin resistence and cancer	no
2	APOM	55937	none		apolipoprotein M involved in lipid transport; associated with RA	no
3	ATXN2L	11273	none		unknown function;associated with a complex group of neurodegenerative disorders.	no
4	BAG6	7017	none		BCL2-associated athanogene 6; located within HLA class III;implicated in the control of apoptosis	no
5	BSN	8927	none		bassoon presynaptic cytomatrix protein; involved in the organization of the cytomatrix at the nerve terminals active zone (CAZ) which regulates neurotransmitter release and in the formation of the retinal photoreceptor ribbon synapse; regulates neurotransmitter release from a subset of brain glutamatergic synapses.	no
6	C2	717	hsa04610	Complement and coagulation cascades	complement component 2; serine endopeptidase; associated with Psoriasis, T1D, Macular degeneration, SLE, carcinoma	yes
			hsa05133	Pertussis		
			hsa05150	Staphylococcus aureus infection		
			hsa05322	Systemic lupus erythematosus		
7	CCDC88B	283234	none		coiled-coil domain containing 88B; function not known	no
8	CCR6	1235	hsa04060	Cytokine-cytokine receptor interaction	G Protein-Coupled Receptors for cytokines; important for B-lineage maturation and antigen-driven B-cell differentiation, and it may regulate the migration and recruitment of dentritic and T cells during inflammatory and immunological responses.	yes
			hsa04062	Chemokine signaling pathway		
9	CTLA4	1493	hsa04514	Cell adhesion molecules (CAMs)		yes
			hsa04660	T cell receptor signaling pathway	immunoglobilin superfamily, encodes a protein which transmits an inhibitory signal to T cells; associated with RA, T1D, Graves disease, Hashimoto thyroiditis, celiac disease, SLE, thyroid-associated orbitopathy, alopecia and other autoimmune diseases	
			hsa05320	Autoimmune thyroid disease		
			hsa05323	Rheumatoid arthritis		
			H00081	Hashimoto's thyroiditis		
			H00082	Graves' disease		
			H00083	Allograft rejection		
			H00408	Type I diabetes mellitus		
10	CD40		hsa04060	Cytokine-cytokine receptor interaction	CD40 molecule, TNF receptor superfamily member 5; mainly is NFkB and MARK dependant, very small involvment in Jak-STAT for INF autokrine activiation loop; mainly is NFkB and MARK dependant, very small involvment in Jak-STAT for INF autokrine activiation loop; GWAS associated RA, MS etc., liked with ALS, Alzheimer and hyper IgM immunodeficiency	yes
			hsa04064	NF-kappa B signaling pathway (noncanonical)		
			hsa04514	Cell adhesion molecules (CAMs)		
			hsa04620	Toll-like receptor signaling pathway		
			hsa04672	Intestinal immune network for IgA production		
			hsa05144	Malaria		
			hsa05145	Toxoplasmosis		
			hsa05166	HTLV-I infection		
			hsa05169	Epstein-Barr virus infection		

			hsa05202	Transcriptional misregulation in cancer		
			hsa05310	Asthma		
			hsa05320	Autoimmune thyroid disease		
			hsa05322	Systemic lupus erythematosus		
			hsa05330	Allograft rejection		
			hsa05340	Primary immunodeficiency		
			hsa05416	Viral myocarditis		
			diseases	Immunodef, Hyper IgM syndrom		
11	CXCR1&2	3579	hsa04060	Cytokine-cytokine receptor interaction		yes
			hsa04062	Chemokine signaling pathway	a receptor for interleukin 8; linked with melanoma developemnt	
			hsa04144	Endocytosis		
			hsa05120	Epithelial cell signaling in Helicobacter pylori infection		
12	ERAP2	64167	hsa04612	Antigen processing and presentation	endoplasmic reticulum aminopeptidase 2; hydrolyzes N-terminal aa of peptide substrates; (MHC) class I molecules rely on aminopeptidases (ERAP1,2 etc.) to trim precursors to antigenic peptides in the endoplasmic reticulum (ER) following cleavage in the cytoplasm	yes
13	GNA12	2768	hsa04010	MAPK signaling pathway		no
			hsa04022	cGMP-PKG signaling pathway	guanine nucleotide binding protein (G protein) alpha 12; G protein; GTP-binding proteins; not in TNF trigered subpath; no known associations	
			hsa04270	Vascular smooth muscle contraction		
			hsa04730	Long-term depression		
			hsa04810	Regulation of actin cytoskeleton		
			hsa05200	Pathways in cancer		
14	GPSM3	63940	none		G-protein signaling modulator 3; associated with Ps and SLE	no
15	HORMAD2	150280	none		HORMA domain containing 2; assiciated with CD, IBD, nephropathy	no
16	IL10	3586	hsa04060	Cytokine-cytokine receptor interaction	IL10 is often anti-immflamatory as IL6 and it is acting via Jak-STAT, no NFkB; has pleiotropic effects in immunoregulation and inflammation	yes
			hsa04068	FoxO signaling pathway	via Jak-STAT first and then to Foxo via STAT3 (same as IL6); an essential immunoregulator linked with infection and RA	
			hsa04630	Jak-STAT signaling pathway		
			hsa04660	T cell receptor signaling pathway	IL10 is a product of this pathway	
			hsa04672	Intestinal immune network for IgA production	Act on B cell same as IL6 (IL6 is not IL10 family but hematopoetins); not included in TNFpathways;	
			hsa05133	Pertussis	IL10 and TNF are only products of this pathway	
			hsa05140	Leishmaniasis	product of this pathway	
			hsa05142	Chagas disease (American trypanosomiasis)	product of this pathway (same as IL6)	
			hsa05143	African trypanosomiasis	product of this pathway	
			hsa05144	Malaria	product of this pathway	
			hsa05145	Toxoplasmosis	acts to IL10R and then via Jak-STAT	
			hsa05146	Amoebiasis	product	
			hsa05150	Staphylococcus aureus infection		
			hsa05152	Tuberculosis	real sugnaling! Acts on cathepsins directly, without Jak_STAT (Cathepsins S,L are proteases in lysosome, phagosome) and inhibits antigen presentaion in Antigen P&P for MHC II to CD4 T! TNF has also real sugnaling to inhibit apoptosis in TB and maybe enchances? Antigen P&P but for MHC I antigens and CD8 T cells	
			hsa05169	Epstein-Barr virus infection	real signaling, acts on IL10R and via Jak-STAT to INFgamma, also a product immuno modulator	
			hsa05310	Asthma	IL10 acts apon mast cells, TNF is product that acts on lung epithel	
			hsa05320	Autoimmune thyroid disease		
			hsa05321	Inflammatory bowel disease (IBD)	real signaling to INF via Jak-STAT and product as well as immuno modulator	

			hsa05322	Systemic lupus erythematosus		
			hsa05330	Allograft rejection		
17	IRF1	3659	hsa04917	Prolactin signaling pathway	a member of the interferon regulatory transcription factor (IRF) family; play roles in regulating apoptosis and tumor-suppressoion	yes
			hsa05133	Pertussis		
			hsa05160	Hepatitis C (engaged Toll-like R)		
			hsa04620	Toll-like receptor signaling pathway		
18	IRF5	3663	hsa04620	Toll-like receptor signaling pathway	a member of the interferon regulatory transcription factor (IRF) family;	yes
19	LSP1	4046	hsa05152	Tuberculosis	lymphocute specific protein 1 encodes an intracellular F-actin binding protein, lymphocute specific protein 1; extressed on lymphocytes, neutrophils, macrophages; regulate motility of lymphocytes, neutrophils, macrophages	yes
20	NKX2-3	159296	none		NK2 homeobox 3, a member of the NKX family of homeodomain transcription factors; linked with cellular differentiation; associated with UC, CD	no
21	PDE2A		hsa00230	Purine metabolism	phosphodiesterase 2A	no
			hsa04022	cGMP-PKG signaling pathway		
			hsa05032	Morphine addiction		
22	PSORS1C1 & PSORS1C2	170679	none		no annotions	no
23	PBX2				pre-B-cell leukemia homeobox 2	no
24	RBPJ	3516	hsa04330	Notch signaling pathway	recombination signal binding protein for immunoglobulin kappa J region; plays a central role in Notch signaling, a signaling pathway involved in cell-cell communication; represses or activates transcription via the recruitment of chromatin remodeling complexes and acts as a transcriptional activator that activates transcription of Notch target genes.	yes
			hsa05169	Epstein-Barr virus infection		
			hsa05203	Viral carcinogenesis		
25	RNF5		hsa04141	Protein processing in endoplasmic reticulum	E3 ubiquitin-protein ligase RNF5	no
26	SLC22A4	6583	hsa05231	Choline metabolism in cancer	solute carrier family 22 (organic cation/zwitterion transporter), member 4; choline receptor transporter; an organic cation transporter and plasma integral membrane protein partially ATP dependent; associated with neuroblastoma.	no
			H00286	Crohn's disease		
26	SLC22A5	6584	hsa05231	Choline metabolism in cancer	solute carrier family 22 (organic cation/zwitterion transporter), member 5; choline receptor transporter; involved in the active cellular uptake of carnitine; associated with asthma and CD	no
			H00286	Crohn's disease		
			H00525	Disorders of fatty-acid oxidation		
27	SMAD3	4088	hsa04068	FoxO signaling pathway	a transcriptional modulator activated by transforming growth factor-beta and has a role in the regulation of carcinogenesis; involved in the active cellular uptake of carnitine; associated with CD, Asthma, hearth diseases;	no
			hsa04110	Cell cycle		
			hsa04144	Endocytosis		
			hsa04310	Wnt signaling pathway		
			hsa04350	TGF-beta signaling pathway		
			hsa04390	Hippo signaling pathway		
			hsa04520	Adherens junction		
			hsa04550	Signaling pathways regulating pluripotency of stem cells		
			hsa05142	Chagas disease (American trypanosomiasis)		
			hsa05161	Hepatitis B		
			hsa05166	HTLV-I infection		
			hsa05200	Pathways in cancer		
			hsa05210	Colorectal cancer		
			hsa05212	Pancreatic cancer		

			hsa05220	Chronic myeloid leukemia		
			hsa05321	Inflammatory bowel disease (IBD)		
28	SKIV2L		hsa03018	RNA degradation	antiviral helicase SKI2	no
29	TNFRSF14	8764	hsa04060	Cytokine-cytokine receptor interaction	a member of the tumor necrosis factor (TNF) ligand family; a herpesvirus entry mediator, it enables entry of HSV into cells; associated with UC; shown to stimulate the proliferation of T cells, and triggers apoptosis of various tumor cells; also prevent TNFA mediated apoptosis	yes
			hsa05168	Herpes simplex infection		
30	TNPO3	23534	none		transportin 3; associated with RA, Sclerosis, Cirrhosis, Scleroderma, linked to muscular dystrophy	no
31	TNFSF15	9966	hsa04060	Cytokine-cytokine receptor interaction	tumor necrosis factor (ligand) superfamily, member 15; not expressed in either B or T cells; is inducible by TNF and IL-1 alpha, can activate NF-kappaB and MAP kinases, and acts as an autocrine factor to induce apoptosis in endothelial cells and may function as an angiogenesis inhibitor.	yes
			H00286	Crohn's disease		
32	TRAF1	7185	hsa04064	NF-kappa B signaling pathway		yes
			hsa04668	TNF signaling pathway	TNF associated factor 1 is required with TRAF2 for TNF-alpha-mediated activation of MAPK8/JNK and NF-kappaB; mediates the anti-apoptotic signals from TNF receptors; can be induced by EBV. associated only with RA, liked with canacers and infection	
			hsa05168	Herpes simplex infection		
			hsa05169	Epstein-Barr virus infection		
			hsa05200	Pathways in cancer		
			hsa05202	Transcriptional misregulation in cancer		
			hsa05203	Viral carcinogenesis		
			hsa05222	Small cell lung cancer		
33	UBE2L3		hsa04120	Ubiquitin mediated proteolysis	ubiquitin-conjugating enzyme, ubiquitin-protein ligase; The modification of proteins with ubiquitin is an important cellular mechanism for targeting abnormal or short-lived proteins for degradation.	no
			hsa05012	Parkinson's disease		
34	RNF5		hsa04141	Protein processing in endoplasmic reticulum	E3 ubiquitin-protein ligase RNF5	no
35	ZPBP2		none		zona pellucida binding protein 2	no

Note: shaded are genes with Immune system participation

Provided are gene official symbols and IDs, names and IDs of KEGG pathways, gene function, association with immune/inflammatory diseases, or other relevant gene info, and evaluation whether the gene is considered linked directly to the immune system or not.

Supplemental Table 4. Comparison between all pathways: missSNP gene pathways, ncSNP gene pathways and TNF containing pathways

ncSNP pathways	
hsa04060	Cytokine-cytokine receptor interaction
hsa04062	Chemokine signaling pathway
hsa04064	NF-kappa B signaling pathway
hsa04068	FoxO signaling pathway
hsa04144	Endocytosis
hsa04330	Notch signaling pathway
hsa04350	TGF-beta signaling pathway - Homo sapiens (human)
hsa04514	Cell adhesion molecules (CAMs)
hsa04610	Complement and coagulation cascades
hsa04612	Antigen processing and presentation
hsa04620	Toll-like receptor signaling pathway
hsa04630	Jak-STAT signaling pathway
hsa04660	T cell receptor signaling pathway
hsa04668	TNF signaling pathway
hsa04672	Intestinal immune network for IgA production
hsa04917	Prolactin signaling pathway
hsa05012	Parkinson's disease
hsa05120	Epithelial cell signaling in Helicobacter pylori infection
hsa05133	Pertussis
hsa05140	Leishmaniasis
hsa05142	Chagas disease (American trypanosomiasis)
hsa05143	African trypanosomiasis
hsa05144	Malaria
hsa05145	Toxoplasmosis
hsa05146	Amoebiasis
hsa05150	Staphylococcus aureus infection
hsa05152	Tuberculosis
hsa05160	Hepatitis C (engaged Toll-like R)
hsa05166	HTLV-I infection
hsa05168	Herpes simplex infection
hsa05169	Epstein-Barr virus infection
hsa05200	Pathways in cancer
hsa05202	Transcriptional misregulation in cancer
hsa05203	Viral carcinogenesis
hsa05222	Small cell lung cancer
hsa05310	Asthma
hsa05320	Autoimmune thyroid disease
hsa05321	Inflammatory bowel disease (IBD)
hsa05322	Systemic lupus erythematosus
hsa05323	Rheumatoid arthritis
hsa05330	Allograft rejection
hsa05340	Primary immunodeficiency
hsa05416	Viral myocarditis

missSNP pathways	
hsa04060	Cytokine-cytokine receptor interaction
hsa04064	NF-kappa B signaling pathway
hsa04066	HIF-1 signaling pathway
hsa04140	Regulation of autophagy
hsa04145	Phagosome
hsa04151	PI3K-Akt signaling pathway
hsa04330	Notch signaling pathway
hsa04380	Osteoclast differentiation
hsa04611	Platelet activation
hsa04612	Antigen processing and presentation
hsa04621	NOD-like receptor signaling pathway
hsa04630	Jak-STAT signaling pathway
hsa04640	Hematopoietic cell lineage
hsa04660	T cell receptor signaling pathway
hsa04662	B cell receptor signaling pathway
hsa04664	Fc epsilon RI signaling pathway
hsa04666	Fc gamma R-mediated phagocytosis
hsa04668	TNF signaling pathway
hsa04722	Neurotrophin signaling pathway
hsa04919	Thyroid hormone signaling pathway
hsa04920	Adipocytokine signaling pathway
hsa04932	Non-alcoholic fatty liver disease (NAFLD)
hsa05131	Shigellosis
hsa05140	Leishmaniasis
hsa05145	Toxoplasmosis
hsa05150	Staphylococcus aureus infection
hsa05152	Tuberculosis
hsa05160	Hepatitis C
hsa05162	Measles
hsa05164	Influenza A
hsa05168	Herpes simplex infection
hsa05169	Epstein-Barr virus infection
hsa05206	MicroRNAs in cancer
hsa05310	Asthma
hsa05321	Inflammatory bowel disease (IBD)
has05340	Primary immunodeficiency

TNF pathways	
hsa04060	Cytokine-cytokine receptor interaction
hsa04064	NF-kappa B signaling pathway
hsa04150	mTOR signaling pathway - Homo sapiens (human)
hsa04210	Apoptosis
hsa04350	TGF-beta signaling pathway - only probable, not confirmed in KEGG
hsa04380	Osteoclast differentiation
hsa04612	Antigen processing and presentation
hsa04620	Toll-like receptor signaling pathway
hsa04621	NOD-like receptor signaling pathway
hsa04622	RIG-I-like receptor signaling pathway - Homo sapiens (human)
hsa04640	Hematopoietic cell lineage
hsa04650	Natural killer cell mediated cytotoxicity - Homo sapiens (human)
hsa04660	T cell receptor signaling pathway
hsa04664	Fc epsilon RI signaling pathway
hsa04668	TNF signaling pathway
hsa04920	Adipocytokine signaling pathway
hsa04930	Type II diabetes mellitus - Homo sapiens (human)
hsa04932	Non-alcoholic fatty liver disease (NAFLD)
hsa04940	Type I diabetes mellitus - Homo sapiens (human)
hsa05010	Alzheimer's disease - Homo sapiens (human)
hsa05014	Amyotrophic lateral sclerosis (ALS) - Homo sapiens (human)
hsa05133	Pertussis
hsa05134	Legionellosis - Homo sapiens (human)
hsa05140	Leishmaniasis
hsa05142	Chagas disease (American trypanosomiasis)
hsa05143	African trypanosomiasis
hsa05144	Malaria
hsa05145	Toxoplasmosis
hsa05146	Amoebiasis
hsa05152	Tuberculosis
hsa05160	Hepatitis C
hsa05161	Hepatitis B
hsa05164	Influenza A
hsa05166	HTLV-I infection
hsa05168	Herpes simplex infection
hsa05205	Proteoglycans in cancer
hsa05310	Asthma
hsa05321	Inflammatory bowel disease (IBD)
hsa05322	Systemic lupus erythematosus
hsa05323	Rheumatoid arthritis
hsa05330	Allograft rejection
hsa05332	Graft-versus-host disease
hsa05410	Hypertrophic cardiomyopathy (HCM)
hsa05414	Dilated cardiomyopathy

KEGG pathway name and ID is provided for each group of ncSNP genes, missSNP genes and TNF; pathway name underlined by color coded background for easier comparison: yellow common to all three groups, green common to ncSNP genes and TNF, orange for ncSNP genes and missSNP genes; blue common to missSNP genes and TNF.


pathways common to missSNP pathways and ncSNP pathways
pathways common to missSNP pathways, ncSNP pathways and TNF pathways
pathways common to ncSNP pathways and TNF pathways
pathways common to missSNP pathways and TNF pathways

Subpplemental Table 5.

Intersections between TNF signaling pathway and AID GWAS SNP gene pathways AID GWAS SNP analysed by Cytoscape

	Pathways	TNF signaling	NF-kB signaling pathway	NOD2 like signaling	Jak-STAT signaling	Antigen procesing and presenting	B cell R signaling	T cell R signaling	Cytokine-cytokine R interaction	Intestinal immune network for IgA production
1	**TNF signaling**									
2	**NF-kB signaling**	Ubuquitin mediated proteolysis, Apoptosis; TNFAIP3 among 17 members, including NFKBIA, TNF, IKBKG, RIPK1, TRAF1,2, CHUK etc								
3	**NOD like signaling**	Apopotisis, Ubiquitin, MAPK and NFkB signaling; NOD2, TNFAIP3, NFKBIA-NFKB1, TNF etc. 15 genes and 4 processes	TNF, TNFAIP3 TRAF6 NFKBIA , ubiquitin, apoptosis CXCL2 and L8							
4	**Jak-STAT signaling pathway**	AKT3-PIK3R5, CSF5, bacially only AKT and PI3K; Ubuquitin mediated proteolysis, Apoptosis, MAPK signaling, PI3K-Akt signaling pathway are all downstream.	basically NONE! All are downstream of action: BCL2L1 (a product after DNA is a apopotosis regulator); Ubuquitin mediated proteolysis, Apoptosis, Cytok-CytokR	none; apopotosis, MAPK signaling, Ubiquiting mediated proteolysis all downstream						
5	**Antigen procesing and presenting**	TNF	TNF, proteosome, T cell R signaling, K10784	TNF	none					
6	**B cell signaling**	NFKBIA-NFKB1, AKT3-PIK3R5, MAPK2K1-MARK1; FOS; PI3K-Akt signaling pathway, NF-kB signaling pathway, Ubiquitin mediated proteolysis.	NFKBIA SYK LYN BTK ubiqutin mediated proteolysis, calcium signaling	NFKB1- NFKBIA, Ubuquitin proteolysis, MAPK signaling, NFkB signaling	none	none				
7	**T cell signaling**	MAPK1-MAP2K1, AKT3-PIK3R5, NFKB1-NFKBIA; MAPK14, MAP3K8, MAP2K7, MAP3K7, MAP3K14, CSF2, FOS, TNF; Ubiquitin med proteolysis, NFkB signaling, PI3K-Akt signaling, MAPK signaling	TNF NFKBIA AKT PI3K Mapk FOS and many others (20+)	TNF NFKBA NFKB1 MAPK signaling, NFkB signaling, Ubiquitin mediated proteolysis	none	TNF, CD4 and IFNG	many common genes			
8	**Cytokine-cytokine R interaction***	LTA-TNFRSF1B, TNF-TNFRSF1A, LIF, IL6, IL15, IL1B, CSF1, CXCL1, CXCL5,CX3CL1, CXC10, CCL5, CCL2, CCL20, FAS, CSF2 (18 total)	CD40, CD40L TNF TNFRSF1A and a few TNF-TNFR family	TNF and IL6 IL18 IL1R CXCL1,2,8 and CCL2 and 5	IL24 IL22R CSF2 (granulocyte-macrophage colony-stimulating factor)	TNF INFG	none?	none?		
9	**Intestinal immune network for IgA production**	IL15, MAP3K14 (NIK), IL6	TNFRSF13 CD40 BCR TCR signaling etc.	none	none	NA	many genes	many genes	many genes	
10	**Osteoclast differenciation**	TNF, NFkB signaling	TNF MAP3K14 (NIK) NFKBIA TRAF2 NFKB2 CHUK IKBKG SYK IL1R1 TNFRSF11 TNFRSF1A	TNF NFKBIA NFKB1 MAPK1,8,11 NFkB signaling, MAPK signaling	MAPK signaling PI3K-AKT signaling Jak-STAT signaling PIK3R5 IRFS GRB2 AKT3 JAK1 etc.	TNF INFG CREB1	NA	NA	TNF TNFSF11 TNFRSF11A TNFRSF1A IL1A IFNG	
11	**Notch signaling pathway**	none	none	none	none	none	none	none	none	none

* **not a real a pathway, but the coplex collection of cytokines and cytokines receptors**

Supplemental Table 6.

Enriched pathway-based set for missSNP geneset by ConsensusPathDB

uploaded list:	23
mapped entities:	22
enriched pathway-based sets:	78

17 genes (77.3%) from the input list are present in at least one pathway.

pathway name	set size	candidates contained	p-value	q-value	pathway source
JAK-STAT-Core	104	5 (5.0%)	3.53E-07	3.99E-05	Signalink
Jak-STAT signaling pathway - Homo sapiens (human)	156	5 (3.4%)	2.40E-06	0.000135	KEGG
Immune System	1069	9 (0.9%)	5.74E-06	0.000216	Reactome
Signaling by Interleukins	117	4 (3.8%)	1.84E-05	0.000428	Reactome
NOD1/2 Signaling Pathway	35	3 (8.8%)	1.90E-05	0.000428	Reactome
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways	49	3 (6.2%)	5.40E-05	0.00102	Reactome
MAP kinase activation in TLR cascade	59	3 (5.5%)	8.14E-05	0.00108	Reactome
ERK2 activation	10	2 (22.2%)	8.38E-05	0.00108	Reactome
NOD-like receptor signaling pathway	57	3 (5.3%)	9.06E-05	0.00108	KEGG
ERK1 activation	10	2 (20.0%)	0.000105	0.00108	Reactome
Interleukin-6 signaling	11	2 (20.0%)	0.000105	0.00108	Reactome
Inflammatory bowel disease (IBD)	67	3 (4.8%)	0.000122	0.00115	KEGG
ERK activation	13	2 (16.7%)	0.000153	0.00123	Reactome
Cytokine Signaling in Immune system	198	4 (2.2%)	0.000158	0.00123	Reactome
TRAF6 Mediated Induction of proinflammatory cytokines	74	3 (4.3%)	0.000167	0.00123	Reactome
Innate Immune System	607	6 (1.0%)	0.000175	0.00123	Reactome
MyD88 cascade initiated on plasma membrane	87	3 (3.6%)	0.000277	0.00147	Reactome
Toll Like Receptor 10 (TLR10) Cascade	87	3 (3.6%)	0.000277	0.00147	Reactome
Toll Like Receptor 5 (TLR5) Cascade	87	3 (3.6%)	0.000277	0.00147	Reactome
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	88	3 (3.6%)	0.000287	0.00147	Reactome
MyD88 dependent cascade initiated on endosome	90	3 (3.5%)	0.000308	0.00147	Reactome
Toll Like Receptor 7/8 (TLR7/8) Cascade	90	3 (3.5%)	0.000308	0.00147	Reactome
RAF/MAP kinase cascade	19	2 (11.8%)	0.000314	0.00147	Reactome
Toll Like Receptor 9 (TLR9) Cascade	94	3 (3.3%)	0.000352	0.00147	Reactome
Toll Like Receptor TLR1:TLR2 Cascade	97	3 (3.3%)	0.000364	0.00147	Reactome
Toll Like Receptor 2 (TLR2) Cascade	97	3 (3.3%)	0.000364	0.00147	Reactome
MyD88:Mal cascade initiated on plasma membrane	97	3 (3.3%)	0.000364	0.00147	Reactome
Toll Like Receptor TLR6:TLR2 Cascade	97	3 (3.3%)	0.000364	0.00147	Reactome
GRB2 events in EGFR signaling	23	2 (10.0%)	0.000438	0.00156	Reactome
SOS-mediated signalling	23	2 (10.0%)	0.000438	0.00156	Reactome
TRIF-mediated TLR3/TLR4 signaling	105	3 (3.0%)	0.00048	0.00156	Reactome
MyD88-independent cascade	105	3 (3.0%)	0.00048	0.00156	Reactome
Toll Like Receptor 3 (TLR3) Cascade	105	3 (3.0%)	0.00048	0.00156	Reactome
SHC-mediated signalling	24	2 (9.5%)	0.000484	0.00156	Reactome
SHC1 events in EGFR signaling	24	2 (9.5%)	0.000484	0.00156	Reactome
Cytokine-cytokine receptor interaction	265	4 (1.6%)	0.000515	0.00162	KEGG
Signalling to p38 via RIT and RIN	24	2 (9.1%)	0.000531	0.00162	Reactome
ARMS-mediated activation	25	2 (8.7%)	0.000582	0.00164	Reactome
SHC-related events	27	2 (8.7%)	0.000582	0.00164	Reactome
Canonical NF-kappaB pathway	24	2 (8.7%)	0.000582	0.00164	PID
SHC-related events triggered by IGF1R	27	2 (8.3%)	0.000634	0.00175	Reactome
Activated TLR4 signalling	120	3 (2.7%)	0.000686	0.00177	Reactome
Frs2-mediated activation	27	2 (8.0%)	0.000688	0.00177	Reactome
Signaling by Leptin	29	2 (8.0%)	0.000688	0.00177	Reactome
Prolonged ERK activation events	29	2 (7.4%)	0.000804	0.00197	Reactome
SHC1 events in ERBB4 signaling	30	2 (7.4%)	0.000804	0.00197	Reactome
Toll Like Receptor 4 (TLR4) Cascade	129	3 (2.5%)	0.000858	0.00206	Reactome
Canonical NF-kappaB pathway	23	2 (8.7%)	0.000532	0.00207	PID
GRB2 events in ERBB2 signaling	32	2 (6.9%)	0.000928	0.00218	Reactome
Measles - Homo sapiens (human)	134	3 (2.3%)	0.00103	0.00235	KEGG

SHC1 events in ERBB2 signaling	34	2 (6.5%)	0.00106	0.00235	Reactome
Signalling to RAS	34	2 (6.5%)	0.00106	0.00235	Reactome
Signaling by SCF-KIT	149	3 (2.2%)	0.00125	0.00272	Reactome
VEGFR2 mediated cell proliferation	37	2 (5.9%)	0.00128	0.00272	Reactome
Toll-Like Receptors Cascades	149	3 (2.1%)	0.00133	0.00278	Reactome
IL23-mediated signaling events	38	2 (5.7%)	0.00135	0.00278	PID
NOD pathway	39	2 (5.1%)	0.00168	0.00339	Wikipathways
IL-6 signaling pathway	43	2 (4.9%)	0.00185	0.00361	Wikipathways
Signalling to ERKs	44	2 (4.9%)	0.00185	0.00361	Reactome
FRS2-mediated cascade	46	2 (4.7%)	0.00204	0.00384	Reactome
Interleukin-2 signaling	51	2 (4.7%)	0.00204	0.00384	Reactome
IL6-mediated signaling events	49	2 (4.4%)	0.00223	0.00413	PID
Tuberculosis - Homo sapiens (human)	179	3 (1.7%)	0.00234	0.00427	KEGG
Signaling events mediated by PTP1B	55	2 (4.1%)	0.00264	0.00473	PID
Epstein-Barr virus infection - Homo sapiens (human)	202	3 (1.5%)	0.00344	0.00607	KEGG
IL4	64	2 (3.3%)	0.00393	0.00683	NetPath
NCAM signaling for neurite out-growth	75	2 (3.0%)	0.00488	0.00835	Reactome
IL6	77	2 (2.8%)	0.00561	0.00947	NetPath
Allograft Rejection	80	2 (2.5%)	0.00689	0.0114	Wikipathways
Hematopoietic cell lineage - Homo sapiens (human)	87	2 (2.4%)	0.00723	0.0117	KEGG
FCERI mediated MAPK activation	88	2 (2.4%)	0.00723	0.0117	Reactome
IRS-mediated signalling	89	2 (2.4%)	0.0074	0.0118	Reactome
IRS-related events	92	2 (2.4%)	0.00775	0.0118	Reactome
TNF alpha Signaling Pathway	87	2 (2.4%)	0.00775	0.0118	Wikipathways
NF-kappa B signaling pathway - Homo sapiens (human)	91	2 (2.3%)	0.00828	0.0119	KEGG
IRS-related events triggered by IGF1R	93	2 (2.3%)	0.0081	0.0122	Reactome
Insulin receptor signalling cascade	96	2 (2.2%)	0.00847	0.0125	Reactome
IGF1R signaling cascade	96	2 (2.2%)	0.00865	0.0125	Reactome

Number of queried genes entered for enrichment analyses and number of genes mapped in pathways; number of enriched pathway sets; pathways names with the number of genes in each pathway and percentage of queried genes for each pathway along with number of genes in a gene set for each pathway; p values, after fdr and Bonferroni corrections.

Supplemental Table 7. Enriched pathway-based set for ncSNP geneset by ConsensusPath DB

uploaded list:	23
mapped entities:	23
enriched pathway-based sets:	26

14 genes (60.9%) from the input list are present in at least one pathway.

pathway name	set size	candidates contained	p-value	q-value	pathway source
Chemokine receptors bind chemokines	60	3 (5.3%)	4.91E-05	0.00101	Reactome
Peptide GPCRs	72	3 (5.3%)	4.91E-05	0.00101	Wikipathways
Pertussis - Homo sapiens (human)	75	3 (4.0%)	0.000112	0.00153	KEGG
Cytokine-cytokine receptor interaction - Homo sapiens (human)	265	4 (1.6%)	0.000229	0.00234	KEGG
Senescence and Autophagy	105	3 (2.9%)	0.000303	0.00248	Wikipathways
the information processing pathway at the ifn beta enhancer	30	2 (6.9%)	0.000624	0.00426	BioCarta
Chemokine signaling pathway - Homo sapiens (human)	189	3 (1.6%)	0.00153	0.00784	KEGG
Peptide ligand-binding receptors	198	3 (1.6%)	0.00153	0.00784	Reactome
Autoimmune thyroid disease - Homo sapiens (human)	54	2 (3.8%)	0.002	0.00827	KEGG
Staphylococcus aureus infection - Homo sapiens (human)	57	2 (3.8%)	0.00208	0.00827	KEGG
Endocytosis - Homo sapiens (human)	213	3 (1.4%)	0.00227	0.00827	KEGG
G alpha (i) signalling events	242	3 (1.4%)	0.00259	0.00827	Reactome
GPCRs, Class A Rhodopsin-like	259	3 (1.4%)	0.00262	0.00827	Wikipathways
Inflammatory bowel disease (IBD) - Homo sapiens (human)	67	2 (3.2%)	0.00293	0.00857	KEGG
Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human)	68	2 (3.0%)	0.0033	0.00872	KEGG
Regulation of Telomerase	71	2 (2.9%)	0.0034	0.00872	PID
Regulation of nuclear SMAD2/3 signaling	79	2 (2.6%)	0.00423	0.01	PID
Allograft Rejection	80	2 (2.5%)	0.00468	0.01	Wikipathways
Apoptosis	84	2 (2.5%)	0.00479	0.01	Wikipathways
GPCR downstream signaling	986	5 (0.5%)	0.0049	0.01	Reactome
GPCRs, Other	98	2 (2.3%)	0.00539	0.0105	Wikipathways
Class A/1 (Rhodopsin-like receptors)	325	3 (1.0%)	0.00569	0.0106	Reactome
Signal Transduction	2077	7 (0.4%)	0.0069	0.0123	Reactome
T cell receptor signaling pathway - Homo sapiens (human)	104	2 (2.0%)	0.00736	0.0125	KEGG
Chagas disease (American trypanosomiasis) - Homo sapiens (human)	104	2 (1.9%)	0.00764	0.0125	KEGG
Signaling by GPCR	1108	5 (0.5%)	0.00811	0.0128	Reactome

Number of queried genes entered for enrichment analyses and number of genes mapped in pathways; number of enriched pathway sets; pathways names with the number of genes in each pathway and percentage of queried genes for each pathway along with number of genes in a gene set for each pathway; p values, after fdr and Bonferroni corrections.

Supplemental Table 8. Enriched pathway-based sets for allSNP set

uploaded list:	56
mapped entities:	54
enriched pathway-based sets:	85

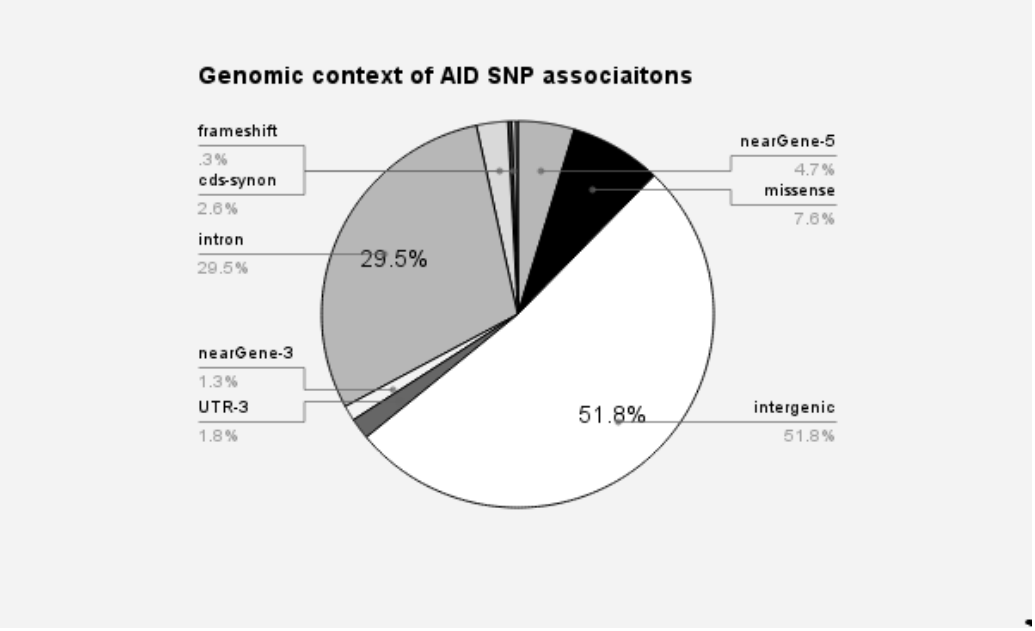
31 genes (68.9%) from the input list are present in at least one pathway

pathway name	set size	candidates contained	p-value	q-value	pathway source
JAK-STAT-Core	104	6 (6.0%)	3.36E-07	2.87E-05	Signalink
Cytokine-cytokine receptor interaction - Homo sapiens (human)	265	8 (3.2%)	3.52E-07	2.87E-05	KEGG
Inflammatory bowel disease (IBD) - Homo sapiens (human)	67	5 (7.9%)	8.80E-07	4.78E-05	KEGG
Immune System	1069	13 (1.3%)	1.38E-06	5.63E-05	Reactome
Jak-STAT signaling pathway - Homo sapiens (human)	156	6 (4.1%)	3.24E-06	0.000106	KEGG
Senescence and Autophagy	105	5 (4.8%)	1.11E-05	0.000302	Wikipathways
Allograft Rejection	80	4 (5.0%)	7.62E-05	0.00177	Wikipathways
NOD1/2 Signaling Pathway	35	3 (8.8%)	0.000122	0.00248	Reactome
Innate Immune System	607	8 (1.4%)	0.000174	0.00312	Reactome
IL-6 signaling pathway	43	3 (7.3%)	0.000214	0.00312	Wikipathways
Signaling by Interleukins	117	4 (3.8%)	0.000219	0.00312	Reactome
Epstein-Barr virus infection - Homo sapiens (human)	202	5 (2.5%)	0.00023	0.00312	KEGG
IL6-mediated signaling events	49	3 (6.7%)	0.000282	0.00332	PID
ERK2 activation	10	2 (22.2%)	0.000285	0.00332	Reactome
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) sign	49	3 (6.2%)	0.000342	0.00341	Reactome
ERK1 activation	10	2 (20.0%)	0.000355	0.00341	Reactome
Interleukin-6 signaling	11	2 (20.0%)	0.000355	0.00341	Reactome
Staphylococcus aureus infection - Homo sapiens (human)	57	3 (5.7%)	0.000459	0.004	KEGG
MAP kinase activation in TLR cascade	59	3 (5.5%)	0.000512	0.004	Reactome
ERK activation	13	2 (16.7%)	0.00052	0.004	Reactome
Chemokine receptors bind chemokines	60	3 (5.3%)	0.000568	0.004	Reactome
Peptide GPCRs	72	3 (5.3%)	0.000568	0.004	Wikipathways
NOD-like receptor signaling pathway - Homo sapiens (human)	57	3 (5.3%)	0.000568	0.004	KEGG
Notch-HLH transcription pathway	13	2 (15.4%)	0.000613	0.004	Reactome
NICD traffics to nucleus	13	2 (15.4%)	0.000613	0.004	Reactome
TRAF6 Mediated Induction of proinflammatory cytokines	74	3 (4.3%)	0.00104	0.00641	Reactome
RAF/MAP kinase cascade	19	2 (11.8%)	0.00106	0.00641	Reactome
Pertussis - Homo sapiens (human)	75	3 (4.0%)	0.00127	0.00706	KEGG
Tuberculosis - Homo sapiens (human)	179	4 (2.3%)	0.00144	0.00706	KEGG
GRB2 events in EGFR signaling	23	2 (10.0%)	0.00147	0.00706	Reactome
SOS-mediated signalling	23	2 (10.0%)	0.00147	0.00706	Reactome
Apoptosis	84	3 (3.7%)	0.00158	0.00706	Wikipathways
SHC-mediated signalling	24	2 (9.5%)	0.00163	0.00706	Reactome
SHC1 events in EGFR signaling	24	2 (9.5%)	0.00163	0.00706	Reactome
MyD88 cascade initiated on plasma membrane	87	3 (3.6%)	0.0017	0.00706	Reactome
Toll Like Receptor 10 (TLR10) Cascade	87	3 (3.6%)	0.0017	0.00706	Reactome
Toll Like Receptor 5 (TLR5) Cascade	87	3 (3.6%)	0.0017	0.00706	Reactome
Cytokine Signaling in Immune system	198	4 (2.2%)	0.00174	0.00706	Reactome
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activa	88	3 (3.6%)	0.00176	0.00706	Reactome
Signalling to p38 via RIT and RIN	24	2 (9.1%)	0.00179	0.00706	Reactome
MyD88 dependent cascade initiated on endosome	90	3 (3.5%)	0.00188	0.00706	Reactome
Toll Like Receptor 7/8 (TLR7/8) Cascade	90	3 (3.5%)	0.00188	0.00706	Reactome
ARMS-mediated activation	25	2 (8.7%)	0.00195	0.00706	Reactome
SHC-related events	27	2 (8.7%)	0.00195	0.00706	Reactome
Canonical NF-kappaB pathway	24	2 (8.7%)	0.00195	0.00706	PID
SHC-related events triggered by IGF1R	27	2 (8.3%)	0.00213	0.00706	Reactome
Toll Like Receptor 9 (TLR9) Cascade	94	3 (3.3%)	0.00214	0.00706	Reactome
Toll Like Receptor TLR1:TLR2 Cascade	97	3 (3.3%)	0.00221	0.00706	Reactome
Toll Like Receptor 2 (TLR2) Cascade	97	3 (3.3%)	0.00221	0.00706	Reactome
MyD88:Mal cascade initiated on plasma membrane	97	3 (3.3%)	0.00221	0.00706	Reactome
Toll Like Receptor TLR6:TLR2 Cascade	97	3 (3.3%)	0.00221	0.00706	Reactome

Frs2-mediated activation	27	2 (8.0%)	0.00231	0.0071	Reactome
Signaling by Leptin	29	2 (8.0%)	0.00231	0.0071	Reactome
Prolonged ERK activation events	29	2 (7.4%)	0.00269	0.00783	Reactome
SHC1 events in ERBB4 signaling	30	2 (7.4%)	0.00269	0.00783	Reactome
Cytokines and Inflammatory Response	29	2 (7.4%)	0.00269	0.00783	Wikipathways
TRIF-mediated TLR3/TLR4 signaling	105	3 (3.0%)	0.00289	0.00798	Reactome
MyD88-independent cascade	105	3 (3.0%)	0.00289	0.00798	Reactome
Toll Like Receptor 3 (TLR3) Cascade	105	3 (3.0%)	0.00289	0.00798	Reactome
T cell receptor signaling pathway - Homo sapiens (human)	104	3 (3.0%)	0.00297	0.00807	KEGG
GRB2 events in ERBB2 signaling	32	2 (6.9%)	0.0031	0.00815	Reactome
the information processing pathway at the ifn beta enhancer	30	2 (6.9%)	0.0031	0.00815	BioCarta
Asthma - Homo sapiens (human)	32	2 (6.7%)	0.00332	0.00858	KEGG
SHC1 events in ERBB2 signaling	34	2 (6.5%)	0.00354	0.00887	Reactome
Signalling to RAS	34	2 (6.5%)	0.00354	0.00887	Reactome
IL12 signaling mediated by STAT4	33	2 (6.2%)	0.00377	0.0093	PID
Activated TLR4 signalling	120	3 (2.7%)	0.00408	0.00992	Reactome
VEGFR2 mediated cell proliferation	37	2 (5.9%)	0.00425	0.0102	Reactome
IL23-mediated signaling events	38	2 (5.7%)	0.00449	0.0106	PID
Toll Like Receptor 4 (TLR4) Cascade	129	3 (2.5%)	0.00505	0.0118	Reactome
NOD pathway	39	2 (5.1%)	0.00556	0.0128	Wikipathways
FoxO signaling pathway - Homo sapiens (human)	134	3 (2.4%)	0.00565	0.0128	KEGG
Measles - Homo sapiens (human)	134	3 (2.3%)	0.00603	0.0135	KEGG
Signalling to ERKs	44	2 (4.9%)	0.00613	0.0135	Reactome
Systemic lupus erythematosus - Homo sapiens (human)	136	3 (2.3%)	0.00642	0.014	KEGG
FRS2-mediated cascade	46	2 (4.7%)	0.00672	0.0142	Reactome
Interleukin-2 signaling	51	2 (4.7%)	0.00672	0.0142	Reactome
Differentiation Pathway	44	2 (4.5%)	0.00703	0.0147	Wikipathways
Signaling by SCF-KIT	149	3 (2.2%)	0.00726	0.015	Reactome
Notch Signaling Pathway	45	2 (4.4%)	0.00735	0.015	Wikipathways
Toll-Like Receptors Cascades	149	3 (2.1%)	0.00769	0.0155	Reactome
Notch signaling pathway - Homo sapiens (human)	48	2 (4.2%)	0.00833	0.0165	KEGG
Signaling events mediated by PTP1B	55	2 (4.1%)	0.00866	0.017	PID
SIDS Susceptibility Pathways	156	3 (1.9%)	0.0096	0.0186	Wikipathways
Autoimmune thyroid disease - Homo sapiens (human)	54	2 (3.8%)	0.00972	0.0186	KEGG

Number of queried genes entered for enrichment analyses and number of genes mapped in pathways; number of enriched pathway sets; pathways names with the number of genes in each pathway and percentage of queried genes for each pathway along with number of genes in a gene set for each pathway; p values, after fdr and Bonferroni corrections.

Supplemental Figure 1. Genomic context distribution of AID GWAS SNPs



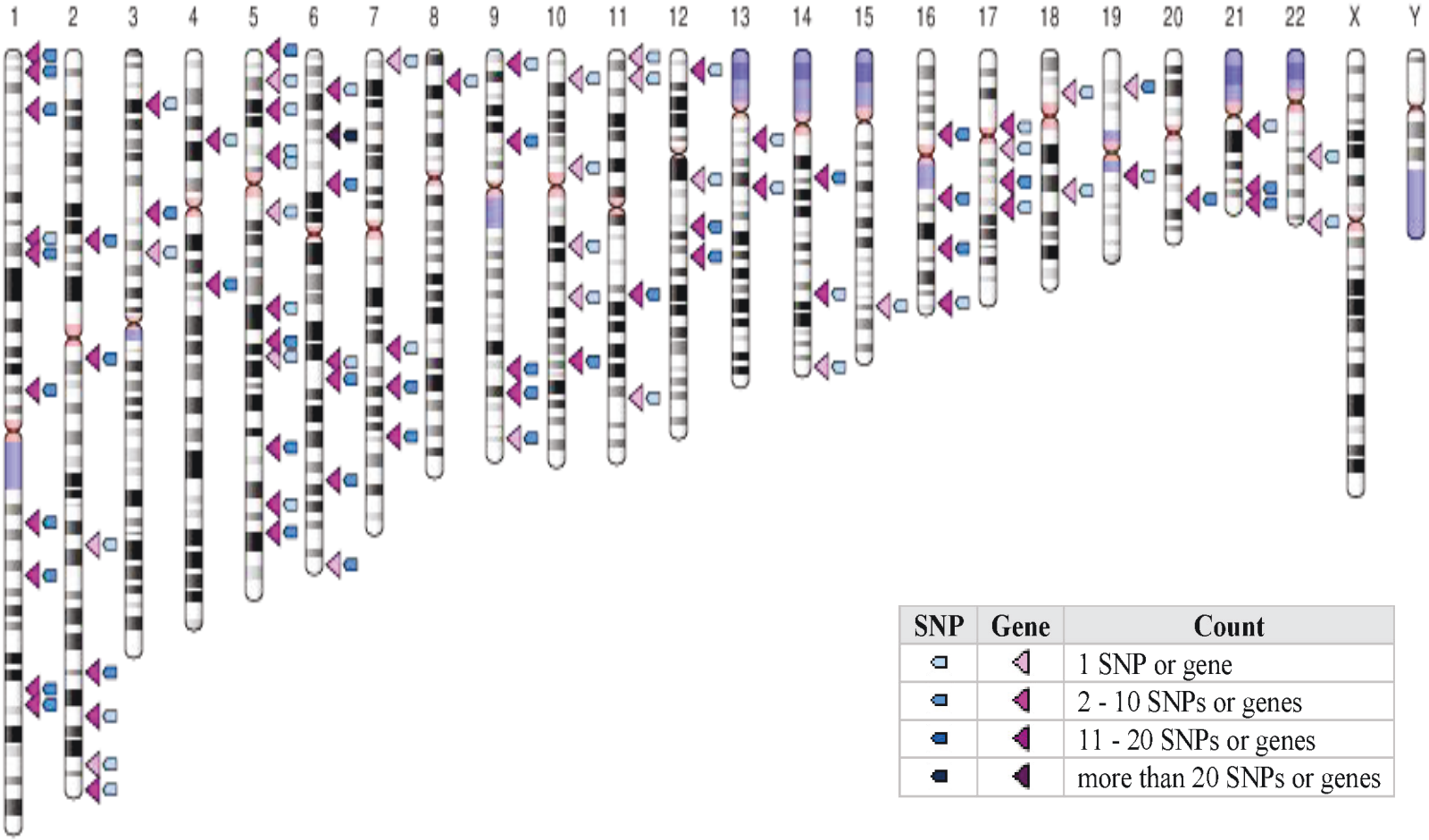
Caption:

AID SNPs were retrieved from NHGRI Catalog by the end of November 2014.

Over 50% of all AID SNPs are intergenic, 30% are intronic and only 7,5% are missense; 3% are UTR-3 and UTR-5 SNPs, 5% are nearGene-5 and 1% are nearGene-3; coding synonymous SNPs represent only 2,5%. There are only 2 frameshift SNP among the retrieved AID GWAS associations.

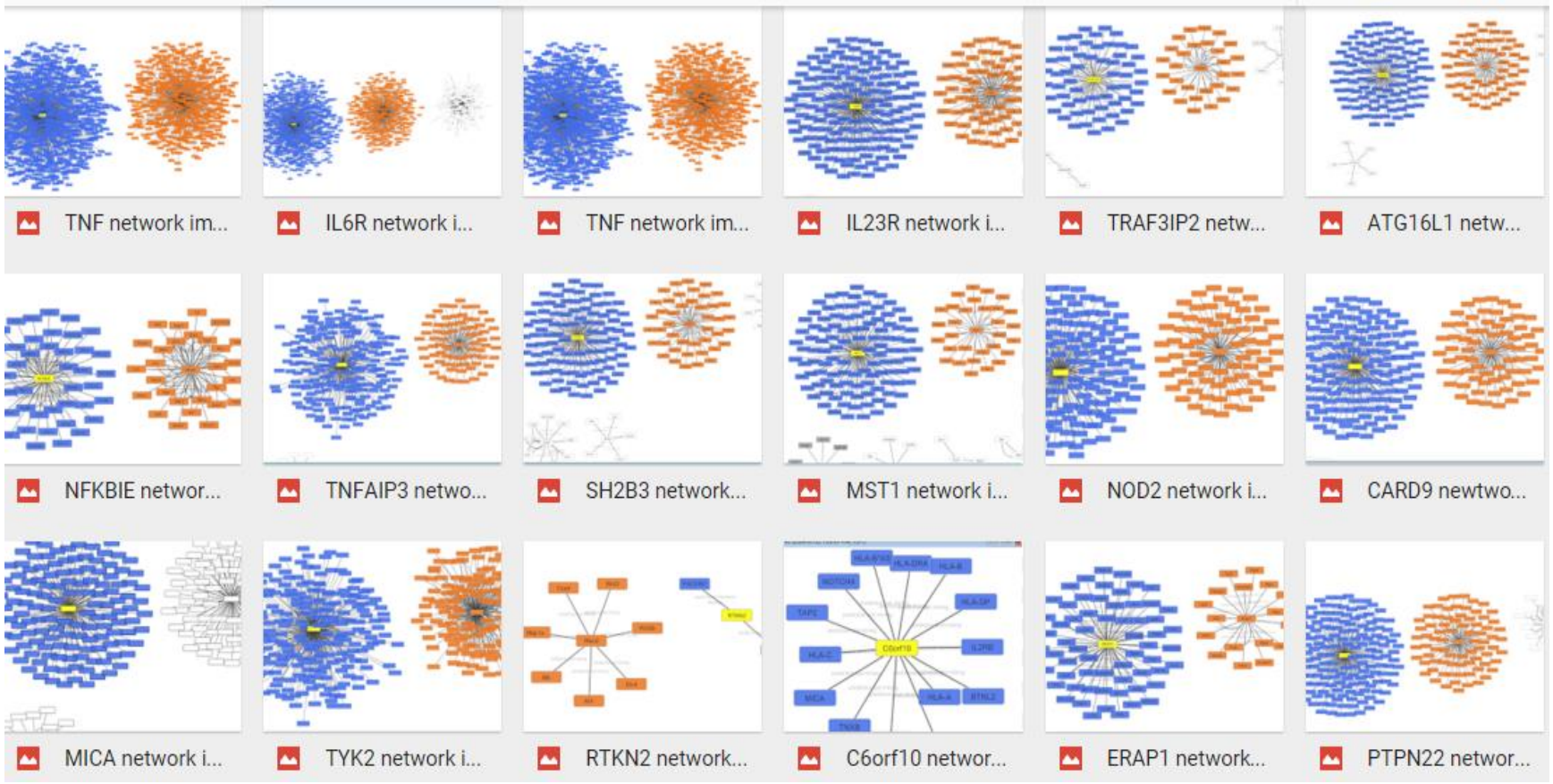
Ideogram of the SNP associations published since the beginning of the GWAS era for the seven autoimmune diseases under our study (AID)

383 SNPs and over 350 genes over 22 chromosomes from all association results.



SNP	Gene	Count
		1 SNP or gene
		2 - 10 SNPs or genes
		11 - 20 SNPs or genes
		more than 20 SNPs or genes

Supplemental Figure 3. Network images for missSNP harboring genes/proteins and TNF



The networks constructed using Cytoscape for the missSNP genes and TNF are presented as yellow colored nodes: all other nodes are color coded for a species (blue for human, red for mouse); nodes represent genes/proteins; edges represent any type of interactions between two proteins as detected by Cytoscape.