Thesis for the Master's degree in Molecular Biosciences
Main field of study in immunogenetics

Hege Dahlen

CTSL2 as a candidate gene for autoimmune diseases – genetic and functional studies to explore its role in type 1 diabetes and myasthenia gravis

60 study points

Department of Molecular Biosciences
Faculty of mathematics and natural sciences
UNIVERSITY OF OSLO, May 2006

ACKNOWLEDGEMENTS

The work presented in this study was carried out at the Institute of Immunology, Rikshospitalet-Radiumhospitalet Medical Center, in the period between June 2005 and March 2006. Supervision has been provided by Senior Scientist Benedicte A. Lie and PhD-student Marte K. Viken at the Institute of Immunology. Inger Sandlie has been the internal supervisor at the Department of Molecular Biosciences, University of Oslo.

First, I would like to thank Benedicte A. Lie and Erik Thorsby for giving me the opportunity to perform my work at the Institute of Immunology. Benedicte; your expertise has been invaluable. Your dedicated support has been highly appreciated, and not to forget your everyday good spirit. I also would like to thank Marte K. Viken for excellent supervision throughout my work. You have always had time for questions and discussion, and your guidance and enthusiasm has been highly appreciated. Don't you ever stop baking those cakes! It has been a pleasure to be part of the Immunogenetics group at the Institute of Immunology, both socially and intellectually. Every single one of you makes a difference.

Thanks to my family for always supporting me, and my friends for still being there… Last, but not least, my dear Jørund. Thank you for believing in me and encouraging me. Your insight into my work means a lot to me, but most of all; thank you for being who you are!

Oslo, May 2006
Hege Dahlen

TABLE OF CONTENTS

ABBREVIATIONS

µL	Microliter
µM	Micromolar
A	Adenine
AChR	Acetylcholine receptor
AFBAC	Affected family-based controls
AIRE	Autoimmune regulator
APC	Antigen presenting cells
B2M	Beta-2-microglobulin
BLAST	Basic Local Alignment Search Tool
bp	Base pair
C	Cytosine
CD	Cluster of differentiation
cDNA	complementary DNA
CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance regulator
CI	Confidence intervals
CLIP	Class II-associated invariant-chain peptide
cTEC	Cortical thymic epithelial cells
CTLA4	Cytotoxic T lymphocyte-associated antigen 4
CTSL2	Cathepsin L2
dbSNP	database SNP
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
E.coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EOMG	Early onset myasthenia gravis
G	Guanine
GR	Genotype risk
HH	Hereditary haemocromatosis
HLA	Human leukocyte antigen
HWE	Hardy-Weinberg equilibrium
IBD	Identical by decent
Ii	Invariant chain
INS	Insulin gene
IPTG	Isopropyl thiogalactoside
kb	Kilo base
LD	Linkage disequilibrium
LOMG	Late onset myasthenia gravis
MAF	Minor allele frequency
Mb	Megabase
MG	Myasthenia gravis
$MgCl_2$	Magnesium Chloride
MHC	Major histocompatibility complex
mM	Millimolar
mRNA	messenger RNA
mTEC	Medullary thymic epithelial cells
NFQ	Non fluorescent quencher
ng	Nanogram
NOD	Non-obese diabetic
NT	Non-transmitted allele/haplotype
OR	Odds ratio
PCR	Polymerase chain reaction
PTPN22	Protein tyrosine phosphatase nonreceptor type 22
RA	Rheumatoid arthritis
RACE	Rapid amplification of cDNA ends
RNA	Ribonucleic acid
RT-PCR	Reverse transcriptase PCR
SLE	Systemic lupus erythematosus
SNP	Single nucleotide polymorphisms
T1D	Type 1 diabetes
T	Thymine
T	Transmitted allele/haplotype
T_C	Cytotoxic T cells
TCR	T cell receptor
TDT	Transmission disequilibrium test
T_H	Helper T cells
Tm	Melting temperature
UTR	Untranslated region
VNTR	Variable number of tandem repeats
X-gal	5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

SUMMARY

Autoimmune diseases are the drawback of having a potent immune defence. Type 1 diabetes and myasthenia gravis are examples of such diseases, where disease development is both complex and multifactorial. In general, both genetic and environmental factors contribute, and among genes found to predispose, several are known to be shared between different autoimmune diseases. Examples of such genes are *CTLA4* and *PTPN22*. This candidate gene study aims at delineating whether the cathepsin L2 (*CTSL2)* gene carry genetic variants predisposing to autoimmune diseases. *CTSL2* is highly expressed in cortical thymic epithelial cells and has been shown to be up regulated in patients with myasthenia gravis. *CTSL2* is an analogue to the murine cathepsin L gene, which previously has been suggested to play a role in development of diabetes in non-obese diabetic mice.

Four single nucleotide polymorphisms (SNP) and two microsatellites covering *CTSL2* were genotyped in 429 Norwegian type 1 diabetes trio families, in addition to 83 German myasthenia gravis patients and 244 German controls. A haplotype consisting of two SNPs (rs10739289 and rs7875800) and one microsatellite (D9S971) was found to be associated with type 1 diabetes (34 transmitted vs. 13 non-transmitted, p=0.002). Another SNP, rs4743056, together with the D9S971 microsatellite, showed association with myasthenia gravis (14.3% in cases vs. 7.5% in controls, p=0.02), and an even stronger association with the early onset subgroup of myasthenia gravis (25% in cases vs. 7.5% in controls, p=0.007). Given the limited polymorphisms studied, and the associations found with different markers in the two diseases, the primary association has probably yet to be identified.

Functional analysis of *CTSL2* was also performed. Expression of *CTSL2* in fetal thymic tissue was observed, in addition to confirming expression in adult thymic tissue. The gene was extensively tested for alternative splice variants and several different mRNA transcripts of *CTSL2* were identified. Furthermore, gene expression of *CTSL2* was measured in 42 thymic tissue samples. We investigated whether the associated genotypes could reflect the earlier observed changes in gene expression, however no correlation was observed.

Our data suggest that *CTSL2* could be a susceptibility gene for development of type 1 diabetes and myasthenia gravis, as well as an interesting candidate gene for other autoimmune diseases.

GENERAL INTRODUCTION

Autoimmune diseases and the immune system

Autoimmune diseases arise when our immune system no longer manages to differ between certain self-antigens and foreign pathogens. The loss of self-tolerance leads to an infiltration of T lymphocytes in the tissue/organs involved, and the consequence is a chronic inflammatory state, where extensive tissue and organ damage can take place. This is the drawback of having a potent immune system designed to recognize all possible antigenic structures. Type 1 diabetes (T1D), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE) and myasthenia gravis (MG) are examples of autoimmune diseases. Autoimmune diseases can be organ specific, like T1D, or systemic, like SLE, where the immune response is directed against antigens expressed in several tissue/organs. Most autoimmune diseases are multifactorial, and in general, both genes and environment contributes to disease susceptibility. Since autoimmune diseases affect approximately 3-5% of the general population (Jacobson et al. 1997), a considerable amount of resources are invested into identifying the disease involved genes. The focus of this introduction will be on how genetic studies are used to identify autoimmune predisposing genes, in particular candidate gene studies related to T1D and MG. Therefore, only a brief introduction to the immune system will be given.

The immune system is an impressive and complex network of tissues, cells and molecules defending our body against infectious agents like bacteria, viruses, fungi and parasites (pathogens). All cells participating in this system originate from pluripotent hematopoietic stem cells in the bone marrow. These precursor cells have the ability to continuously divide and give rise to two daughter cells; one that remains a stem cells and one that differentiate further into either a lymphoid or a myeloid progenitor cell. The lymphoid progenitors give rise to B and T lymphocytes, while the myeloid progenitor gives rise to all other cells in the immune system, like erythrocytes, granulocytes, monocytes, macrophages and dendritic cells. The immune system is divided into two defence strategies; the adaptive and the innate immune systems.

The innate immune system, which we are born with, functions as a first line of defence against pathogens. Macrophages, monocytes, granulocytes and natural killer cells are part of this system, where recognition, activation and effector mechanisms are non specific. Through a wide array of cell surface receptors, the innate immune system recognizes common pathogenic constituents on the surface of the pathogens, and subsequently triggers a wide immune response.

Adaptive immunity, on the other hand, is characterized by a highly specific reaction against antigenic structures. In addition to specificity, the adaptive immune system is versatile, has memory, is self-limiting and discriminates between self and non-self (foreign) substances. Especially the latter is important, and autoimmune diseases are the consequence when the adaptive immune system loses its tolerance to self-antigens. Normally, the adaptive immune system eliminates foreign antigens, however in the case where the antigen is self, a sustained response might occur. The adaptive immune response is commonly divided into the humoral (antibodies) and the cellular (T cell mediated) response. B lymphocytes have B cell receptors on their surface, and upon antigen recognition, specific antibodies directed against the antigen are produced. Antigens are, as the word implies, *anti*body *gen*erators. Each antibody has a unique structure that can only bind the corresponding antigen.

T lymphocytes are responsible for cell-mediated immunity, and include both cytotoxic (T_C, CD8+) and helper T cells ($T_{H,}$ CD4+). T lymphocytes are educated in the thymus not to elicit an immune response against self-antigens (see below). However, in autoimmune diseases, such as T1D and MG, self reactive T lymphocytes trigger an immune reaction. T lymphocytes recognize antigens presented by human leukocyte antigen (HLA) class I and class II molecules. HLA class I molecules are found on all nucleated cells, and present endogenous peptides from the cytosol. Invading pathogens, like virus, often use cells in our body as production sites. We then rely on HLA class I molecules to present pathogenic antigens to T_C lymphocytes, which then kill the infected cells. HLA class II molecules are found on the surface of professional antigen presenting cells (APC), like dendritic cells, macrophage, monocytes and B cells. These cells present peptides derived from ingested pathogens from the surrounding environment to T_H cells. This triggers the production of signal molecules that will further promote a T_H1, T_H2 or a regulatory T cell response. A T_H1 response leads to the activation of T_C cells, while a T_H2 response promotes activation of B cells and subsequent antibody production. Finally, there are also regulatory T cells. These prevent autoimmunity by repressing the immune response.

Type 1 diabetes

T1D is characterized by an immune-mediated destruction of the insulin-producing β-cells in the pancreas, and the consequence is insulin deficiency (Tisch and McDevitt 1996). Specific T_C cells infiltrate the pancreas and attack the β-cells, as a consequence of activation by T_H cells. Increased expression of HLA class I and class II molecules are observed in the β-cells, and antibodies against pancreatic components, like insulin and glutamic acid decarboxylase, are detected in sera from patients. Several serious complications like blindness, nephropathy and poor blood circulation make T1D a troublesome disease for the patient, even with today's insulin treatment. Individuals are often diagnosed with T1D at early age, however the disease can occur at any age. In general, the worldwide incidence of T1D is rising, particularly among younger children (0-4 years) (Karvonen et al. 1999; Green and Patterson 2001). The incidence of T1D is geographically variable, with higher incidence rates in the Northern hemisphere. In Norway, the incidence is 22.5/100.000 for children between 0-14 years, and the prevalence is 0.35 for the same group (given as point prevalence); the number of citizens with T1D pr 1000 inhabitants at a given time) (Vaaler et al. 2004).

Evidence for a genetic contribution to T1D comes from family- and twin studies. For monozygotic twins, the T1D concordance rate is reported to be from 30% to 70% (Barnett et al. 1981; Kyvik et al. 1995), while for dizygotic twins it ranges from 2.5% to 11%. In line with this, siblings to a T1D patient have an estimated risk of disease development of 6%, which is significantly higher that the average risk of 0.4% seen in the general population (Spielman et al. 1980). The contribution of environmental factors is also evident, as the risk for T1D among monozygotic twins is less than 100%.

Myasthenia gravis

MG is a T cell dependent antibody-mediated autoimmune disease, where antibodies predominantly are aimed at the acetylcholine receptors (AChR) of the neuromuscular junction. This leads to a blockage of signalling trough the receptor, hence, patients experience muscle weakness and striated muscle fatigue (Engel 1984; Drachman 1994; Vincent et al. 2001). Symptoms are often fluctuating, and the disease becomes life threatening if muscles in the respiratory tract are affected. Both males and females are diagnosed with MG, and the disease can occur at any age. The onset for males is predominantly after the age of 50, while for women between ages 20-30. MG has an annual incidence of 0.2-0.4 pr 100.000 and a prevalence of 8.5-12.5 pr 100.000 (Romi et al. 2005).

MG is a heterogeneous disease, and there are many ways to clinically subgroup the patients. First of all, patients are often grouped according to age of onset; early onset MG (EOMG), before the age of 40, and late onset MG (LOMG), after the age of 40. Secondly, the presence of autoantibodies against AChR is a key element in the diagnosis of MG (Patrick and Lindstrom 1973; Lindstrom et al. 1976). Approximately 85% of the patients have these autoantibodies in their serum (seropositive). In 10-15% of the patients, these antibodies are not detectable (seronegative) (Lindstrom et al. 1976), however there are evidence that also these forms of MG are mediated by autoantibodies (Mossman et al. 1986; Hoch et al. 2001). Another group of seronegative patients acquire MG as a direct consequence of mutations in the genes encoding the AChR (Engel et al. 2003), and will not be discussed here. In addition, antibodies against titin, a muscle protein involved in the elasticity of muscle fibres (Labeit and Kolmerer 1995), are used to group the patients. Furthermore, benign (hyperplasia) and malignant (thymoma) abnormalities of the thymus are often observed (reviewed in Hohlfeld and Wekerle 1994). Hyperplasia describes the situation when thymus is enlarged because of increased amount of B cells, while thymoma is cancer.

The EOMG subgroup predominantly includes patients with AChR antibodies, thymus hyperplasia and has a strong female bias. Anti-titin antibodies are seldom observed, and 65% of all MG patients fall into this group (reviewed in Romi et al. 2005). An important observation is the high frequency of concomitant autoimmune diseases in this group (Thorlacius et al. 1989; Oosterhuis 1997). The LOMG group, on the other hand, includes patients with presence of anti-AChR antibodies (but lower concentrations than in the EOMG subgroup) and often thymus atrophy. The occurrence is equal in males and females, and anti-titin antibodies are observed in approximately 50% of the patients. Thymoma is seen in 15% of the MG patients. These patients can have anti-AChR and anti-titin antibodies, and the age of onset reaches a peak at 50 years. Both males and females are affected. Heterogeneity describes the remaining patients (without thymic abnormalities) (reviewed in Romi et al. 2005).

Genetic studies of autoimmune diseases

Two approaches are commonly applied when searching for susceptibility genes in autoimmune diseases; genome-wide screens and candidate gene studies. Genome-wide screens is used as a tool to identify regions of the human genome that might be linked to or associated with disease, and the regions identified are more likely to contain disease susceptibility variant(s). Candidate gene studies, on the other hand, directly investigate genes of possible disease relevance, e.g. selected based on known or hypothetical involvement in disease pathogenesis.

Genome-wide genetic studies – strategies and achievements

Genome-wide studies can be performed in two different ways; genome-wide linkage and genome-wide association. Both linkage and association studies can be applied on the whole genome or a specific genetic region. Linkage studies are based on the fact that genes positioned close to each other on a chromosome, tend to be inherited together. However, occasionally recombination occurs between these genes. Overall, the closer two genes are located, the lower the recombination rate. Genome-wide linkage studies attempt to identify chromosomal regions that are co-inherited by affected individuals more often than would be expected by chance (Risch 1990a; Risch 1990b; Risch 1990c). In complex diseases, families with more than one affected offspring (sib-pairs) are genotyped, often by microsatellites, and the sharing of inherited alleles is compared between the offspring. Results are given as 0, 1 or 2 IBD (Identical By Decent) within each family, and summed over all families. When no linkage is present, e.g. between markers on different chromosomes or far apart on the same chromosome, Mendelian segregation of alleles are expected; i.e. 25% of siblings share no (0 IBD) parental alleles, 50% share one parental allele (1 IBD) and 25% share both (2 IBD) parental alleles (Suarez 1978). Deviation from this indicates linkage between the disease studied and the chromosomal area surrounding the marker. Genome-wide scans have been performed for T1D (Concannon et al. 2005) and numerous other autoimmune diseases (but none for MG), and interestingly, several of the linked areas are overlapping between the different diseases (Becker et al. 1998). Genome-wide linkage studies have proven useful for studying monogenic diseases, but less efficient for complex disease, where several genes contribute to disease susceptibility. As a matter of fact, none of the performed genome-wide linkage scans for T1D have subsequently led to the discovery of novel disease associated genes. Both HLA and *CTLA4* can be seen in these linkage scans; however both were initially identified through candidate gene studies. In addition, there is often a discrepancy seen between the identified regions when scans are compared.

The regions identified by linkage studies are often large, and a way to narrow down the regions is to perform association analysis. In association studies, the polymorphisms used need to be more densely packed than in genome-wide linkage studies, usually 10-100 times more polymorphisms are genotyped. This is because association relies on the presence of linkage disequilibrium (LD) between one of the tested polymorphisms and the actual disease involved allele(s). LD refers to the situation when two alleles at linked loci appear together as a haplotype more frequently than excepted by chance. On the population level, the surrounding chromosomal area has been subject to recombination over an extended period of time. Affected individuals carry the disease predisposing allele(s) more often than unaffected individuals, however also neighbouring alleles in LD occur at an increased frequency in patients. Considering a polymorphism used in such a study; the closer it is to the actual predisposing allele, the more likely it will be observed at a higher frequency among affected individuals due to LD. As a consequence, association studies can narrow down the regions of interest, and both family and case-control materials can be used (Lander and Schork 1994). In families, the transmission pattern of tested polymorphisms from parents to offspring is studied. When case-control materials are applied, the allele frequencies are compared between a group of patients (cases) and unrelated, but ethnically matched, controls. So far no genome-wide association study has been performed for either T1D or MG, however given the new era of high-throughput genotyping methods, this will probably soon be achieved.

Candidate gene studies – strategies

The identified regions often contain several genes, and further fine mapping association studies, and eventually candidate gene studies, are performed to investigate whether a specific gene is involved in the disease. When performing a candidate gene study statistical methods are used to investigate the possible existence of a correlation between a specific polymorphism and a disease. Candidate gene studies rely on a hypothesis concerning the gene(s) to be studied, and polymorphisms are selected to cover the gene of interest. This is in contrast to linkage studies, were anonymous polymorphisms evenly spaced throughout the genome is genotyped (Tabor et al. 2002).

There are many ways to select a candidate gene. Examples of good candidates are genes with known or hypothetical biological functions related to disease pathogenesis, or genes found to be differentially expressed in patients and controls. If the molecular pathway(s) affected in the disease is known, genes transcribing proteins, signal molecules, receptor molecules, transcription factors etc in this pathway, are also good candidates. Genes involved in the immune system and genes involved in other autoimmune diseases, (e.g. *CTLA4*, *PTPN22*; se below), are of high interest, since several genes are found to predispose to more than one autoimmune disease. On the other hand, current knowledge of disease pathogenesis in not necessarily correct or complete, hence we are likely to miss good candidate genes due to lack of knowledge (Tabor et al. 2002).

It is not necessary to precede a candidate gene study with a genome wide study. Several of today's known susceptibility genes can not be found by genome-wide studies, e.g. the confirmed T1D associated protein tyrosine phosphatase nonreceptor type 22 *(PTPN22)* gene was undetectable by linkage in a scan comprising nearly 1500 families (Concannon et al. 2005). Candidate gene studies are widely used alone, and several disease associated polymorphisms have been identified by this approach. In fact all established predisposing genes for T1D (HLA, *PTPN22*, *INS* and *CTLA4*; see below) were initially detected through candidate gene studies. During the last 30 years, specific variants of more than 50 different genes have been reported to be associated with autoimmune diseases (Ioannidis et al. 2003; Lohmueller et al. 2003). Disease associated alleles can, in some cases, be rare and specific to some populations, however more often they are common variants in the population. This observation led to the common disease-common variant hypothesis (Lander 1996; Risch and Merikangas 1996; Collins et al. 1999).

As previously mentioned, disease susceptibility in complex diseases, like autoimmune diseases, is generally not caused by a single genetic variant. It is usually an intricate cooperation between several susceptibility loci. Furthermore, a specific polymorphism identified in one disease might be involved in other autoimmune diseases, as some genes are thought to be shared. In line with this, clustering of different autoimmune diseases, within families or even the same patient, is often observed (Bias et al. 1986). It seems that genes predisposing to autoimmune diseases roughly can be divided into two classes; those that predisposes for autoimmune diseases in general, i.e. involved in the autoimmunity process, and those only predisposing to a specific disease, maybe directing organ specificity (Becker 1999).

Candidate gene studies in T1D and MG – achievements

For T1D and probably MG, as for many other autoimmune diseases, the most important inherited genetic effects lie within the major histocompatibility complex (MHC), also called the HLA complex (The MHC sequencing consortium 1999). The MHC is located on the short arm of chromosome 6 and is densely packed with genes. More than 40% of these genes are predicted to encode molecules involved in the immune system (The MHC sequencing consortium 1999). The complex was initially divided into the class I, class II and class III region, together covering a distance of 3.6 megabases (Mb). Later studies led to an extension in both directions from the classical MHC, covering a region of 7.6 Mb; hence it was renamed the extended MHC (Horton et al. 2004). Several studies have confirmed genes in this complex to be either positively or negatively associated with T1D, and the main finding is variants of the HLA-DR and -DQ genes in the MHC class II region (Owerbach et al. 1983). In Caucasians, the HLA alleles DQB1*0302-DQA1*0301 (DQ8) and DQB1*0201-DQA1*0501 (DQ2) show the strongest association with T1D (Sheehy et al. 1989), while DQB1*0602-DQA1*0102 (DQ6) has a protective effect even in the presence of autoantibodies (Pugliese et al. 1995). 90% of all T1D patients carry a DQ8 and/or DQ2 containing haplotype, compared to 20% in the general population (Redondo et al. 2001). Individuals with both alleles (heterozygous), have the highest risk of T1D development. The associated DQ8 allele is influenced by which subtype of DRB1*04 (DR4) the individual has; some increases the susceptibility while others reduce the risk (Sheehy et al. 1989; Caillat-Zucman et al. 1992; Cucca et al. 1995; Undlien et al. 1997). The HLA DR and DQ genes can not alone explain the entire genetic predisposition conferred by the MHC, hence other, yet unidentified genes are to be found. Evidence for additional disease associated genes have been found both in the class III/centromeric class I region (reviewed in Lie and Thorsby 2005), as well as in the extended class I region (Lie et al. 1999; Johansson et al. 2003).

Candidate gene studies have also revealed several genes elsewhere in the genome to predispose to T1D. Already identified genes are specific variants of the insulin gene (*INS*, also called *IDDM2*) (Bennett et al. 1995; Undlien et al. 1995), the T_C lymphocyte-associated antigen 4 (*CTLA4*) (Ueda et al. 2003) and *PTPN22* (Bottini et al. 2004; Onengut-Gumuscu et al. 2004; Smyth et al. 2004; Ladner et al. 2005; Qu et al. 2005; Zheng and She 2005; Zhernakova et al. 2005). The insulin gene is located on chromosome 11, and different alleles of a variable number of tandem repeats (VNTR) upstream of the gene are associated with variable expression of *INS* messenger RNA (mRNA) in thymus and pancreas (Pugliese et al. 1997; Vafiadis et al. 1997). The protective variants have been found to increase the insulin mRNA expression in thymus, and thereby possibly promote negative selection of insulin autoreactive T cells (Pugliese et al. 1997; Vafiadis et al. 1997). *CTLA4* is located on chromosome 2, and polymorphisms downstream of the gene have been found to increase the risk of T1D development (Ueda et al. 2003). The *CTLA4* gene encodes a co-stimulatory molecule, which is expressed on the membrane of T cells some days after antigen stimulation. CTLA4 sends inhibitory signals into the T cell, and down regulates the activation when needed. A single nucleotide polymorphism (SNP) downstream of the gene affects the mRNA splicing, and higher levels of a soluble variant is produced. The balance between membrane bound and soluble CTLA4 is changed, which subsequently leads to prolonged activation of T cells. A SNP in the *PTPN22* gene has also been found to be associated with T1D (Bottini et al. 2004). This gene maps to chromosome 1, and is known to down regulate T cell activation (Cohen et al. 1999; Hill et al. 2002). The associated SNP leads to a more potent negative regulator of T lymphocyte activation; hence it is an example of a gain-of-function variant (Vang et al. 2005). Both *CTLA4* and *PTPN22* are also associated with several other autoimmune diseases including SLE (Liu et al. 1998; Kyogoku et al. 2004), RA (Seidl et al. 1998; Begovich et al. 2004; Viken et al. 2005) and MG (Huang et al. 1998; Vandiedonck et al. 2006).

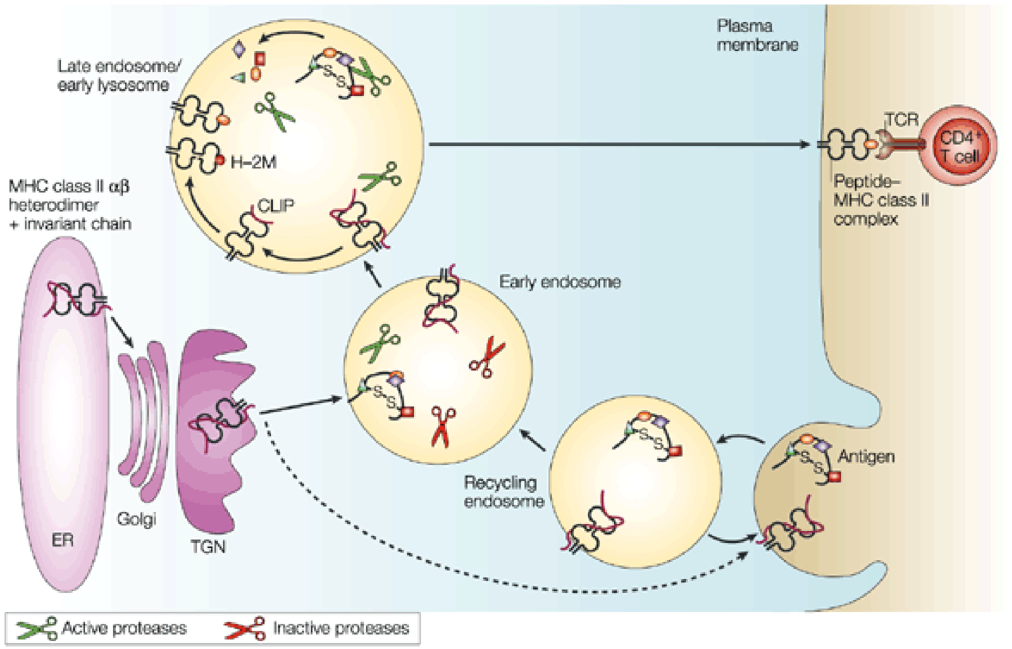
As for T1D, an association has been found between MG and genes in the MHC. The HLA 8.1 haplotype, characterized by HLA-A*01-B*08-DRB1*03-DQA1*0501-DQB1*0201, is associated with the large subgroup of MG patients that suffer from thymus hyperplasia (Feltkamp et al. 1974; Pirskanen 1976; Dawkins et al. 1983; Carlsson et al. 1990; Spurkland et al. 1991; Hjelmstrom et al. 1998). The association was recently limited to the regions between the HLA-B gene and the tumour necrosis factor or complement genes (Vandiedonck et al. 2004). The 8.1 haplotype is particularly interesting in the sense that it also confers susceptibility to several other autoimmune and immune-related diseases, like T1D, SLE and primary sclerosing cholangitis (Price et al. 1999; Candore et al. 2002). This haplotype is also the most common HLA haplotype in the Norwegian population. At one point in history this haplotype might have been advantageous, and it has been speculated that it might have been protective against tuberculosis (Correa et al. 2005). Interestingly, other subgroups of MG show association with different HLA haplotypes. The LOMG subgroup of MG patients are weakly associated with HLA-A3, B7, DR2, DR4 (Compston et al. 1980; Carlsson et al. 1990), while anti-titin positive patients have an increased frequency of HLA-DR7 (Giraud et al. 2001).

As mentioned above, both *CTLA4* and *PTPN22* are associated with MG; however it is important to point out that the association found between *CTLA4* and MG was limited to a subgroup of the patients, namely those with thymoma. Because of the high degree of heterogeneity seen in MG patients, it is no surprise that some association are found only within a subgroup of the patients. In addition to genes shared with other autoimmune diseases, several genes have been found to exclusively predispose to MG. Among these, Garchon and colleagues (1994) identified the acetylcholine receptor alpha subunit (CHRNA) gene to confer an increased risk for MG development. Only patients with detectable levels of AChR antibodies were included in the study.

HLA molecules and antigen presentation

As mentioned, genes encoding HLA class I and class II molecules are found to be involved in predisposition to a large number to autoimmune diseases, and for many diseases they represent the strongest genetic risk factors (Thorsby and Lie 2005). HLA class I molecules are encoded by the A, B and C genes of the MHC complex, while HLA class II molecules are encoded by the DRA1, DRB1, DQA1, DQB1, DPA1 and DPB1 genes. All together, 6 class I genes and 6 class II genes are translated into protein. All 12 genes, except DRA1, are highly polymorphic, particularly in the sequence coding for the peptide binding groove, and because of the wide variety of alleles the number of combinations is extensive.

HLA class I and class II molecules are transmembrane glycoproteins with extracellular N-terminal structures that present antigen peptides to T lymphocytes. HLA class II molecules are synthesised in the lumen of the endoplasmatic reticulum, and consist of two polypeptide chains, an α-chain and a β-chain (Figure 1). It is important that no peptide binds the peptide-binding groove before the HLA molecule reaches the endosomes (compartments inside the cell with acidic environment); therefore the HLA class II molecule is co-synthesised with a chaperone protein called the Invariant chain (Ii) which covers the peptide-binding groove. The HLA class II molecule is transported through the endosomal pathway; from endoplasmatic reticulum to the Golgi apparatus, via the trans-Golgi network and further to the endosomal compartments. Extracellular derived antigens are processed in the endosomes by proteolytic degradation into shorter peptides. Some HLA class II molecules are transported via the cell surface, but are rapidly internalised into recycling endosomes. Ii guides the HLA molecule all the way, and when they reach the endosomal compartments, a proteolytic degradation of the Ii by cathepsins takes place until only a small part, CLIP (class II-associated invariant-chain peptide), is left in the peptide-binding groove. Eventually, a second chaperone protein, HLA-DM, degrades CLIP and facilitates the binding of peptides into the empty peptide-binding groove. If a peptide is not stably bound to the class II molecule, HLA-DM catalyzes release of the peptide. The HLA class II molecule, with bound antigen, is then transported to the surface of the APC, which displays the antigen to helper T_H cells in the surrounding environment (Figure 1). When no foreign peptides are present, both HLA class I and class II molecules present self-antigens in the peptide binding groove (reviewed in Cresswell 1998).



Nature Reviews | Immunology

Figure 1 Antigen presentation via HLA class II molecules. The HLA class II molecules assemble in the endoplasmatic reticulum (ER) together with the Invariant chain (Ii). Ii function as a chaperon, guiding the HLA molecule through the endosomal pathway, either directly from the trans-Golgi network (TGN) to early endosomes, or via the plasma membrane. Ii–HLA class II complexes at the cell surface are rapidly internalized into recycling endosomes and then trafficked to the early endosomes. Cysteine cathepsins (proteases) in the endosomes are activated when the endosomes mature, and the cathepsins degrade endogenous endosomal proteins, internalized proteins and Ii. Following Ii cleavage, the HLA class II peptide-binding groove remains occupied by the class-II-associated invariant chain peptide (CLIP), which prevents premature peptide loading. Removal of CLIP and loading of peptides is mediated by HLA-DM. These peptide-HLA class II complexes are then trafficked to the plasma membrane where they present their cargo to the TCRs (T cell receptors) on CD4+ T cells in the surrounding environment. Adapted from Honey and Rudensky 2003.

Thymus

The thymus is one of the most important organs in the immune system. It is located just above the heart, and this is where T lymphocytes are educated through positive and negative selection to not elicit an immune response against self-antigens. Immature T lymphocytes are found in the outer cortex, while mature T lymphocytes, together with macrophages and dendritic cells, are found in the inner medulla. After puberty, the thymus starts to shrink in a process called involution, and its contribution of new T lymphocytes decline. Nonetheless, this does not lead to a reduction in the total number of T lymphocytes in the periphery (Aspinall and Andrew 2000).

In humans, immature T lymphocytes are transported from the bone marrow to the thymus primordium (the earliest recognizable stage of development) already in the eight week of gestation (Owen and Ritter 1969; Haynes and Heinly 1995). Later lymphoid progenitor cells enter the thymus through the vascular system. T lymphocytes undergo several developmental stages in the thymus, and interact with stromal cells, like cTECs, medullar thymic epithelial cells (mTECs) and dendritic cells, throughout the process. The stromal cells express chemokines, and the lymphocytes are known to sequentially express chemokine receptors (reviewed in Takahama 2006). At the double positive stage the T lymphocytes express both CD4 and CD8, in addition to T cell receptors (TCRs), on their surface. In that regard they have the possibility to mature into either CD4+ (T_H) or CD8+ (T_C) T cells. The TCRs interact with peptide-HLA complexes that are expressed on the surface of cTECs and bone marrow derived dendritic cells (Kisielow et al. 1988; Jameson et al. 1995). During positive selection, which takes place in the cortex of the thymus, T cells that are able to recognize HLA class II-self peptide complexes with moderate affinity are selected for (Figure 2a). These T cells receive a survival signal, and can further differentiate into CD4 single positive T cells. T cells reacting with too high affinity towards peptide-HLA complex die by apoptosis (programmed cell death), in a process called negative selection (Figure 2a). Cells not receiving the survival signal also die by apoptosis (Figure 2a). T cells surviving these selection processes migrate to the medulla of the thymus (reviewed in Takahama 2006). During this relocation, the lymphocytes express the CC-chemokine receptor 7, while mTECs secrete the ligands for CC-chemokine receptor 7; CC-chemokine ligand 19 and CC-chemokine ligand 21 (Figure 2b) (Ueno et al. 2004). Before the T lymphocytes are exported, they stay approximately 12 days in the thymic medulla, where further negative selection takes place (Egerton et al. 1990). This selection is important for establishing central tolerance to tissue-specific antigens, and the transcriptional factor autoimmune regulator (AIRE) is involved in this process (Zuklys et al. 2000; Derbinski et al. 2005). AIRE promotes expression of tissue-specific antigens in a subset of mTECs and thymic dendritic cells, and the subsequent presentation of these antigens leads to destruction of self reactive T lymphocytes (Figure 2c) (Anderson et al. 2002; Liston et al. 2003; Gallegos and Bevan 2004). Mutations in this gene have been shown to be responsible for autoimmune polyendocrinopathy syndrome type 1 (APS1), which is a monogenic disorder (Nagamine et al. 1997; Aaltonen et al. 1997). The thymic medulla is also believed to be the site for the production of regulatory T cells, and initially AIRE was thought to be involved, however later reports questioned that observation (reviewed in Su and Anderson 2004). There are still several questions regarding AIRE to be answered, but its vital role in preventing autoimmune diseases is nonetheless interesting.

T cells eventually leaving the thymus have the ability to recognize the body's own HLA class II molecules, but not with such high affinity that they can induce activation. Only 1-3% of the original immature T lymphocytes survive these selection processes (Scollay et al. 1980; Egerton et al. 1990; Goldrath and Bevan 1999). It is believed that similar processes are involved when T_C lymphocytes (CD8+) are educated to recognize HLA class I-peptide complexes (Cresswell 1998).

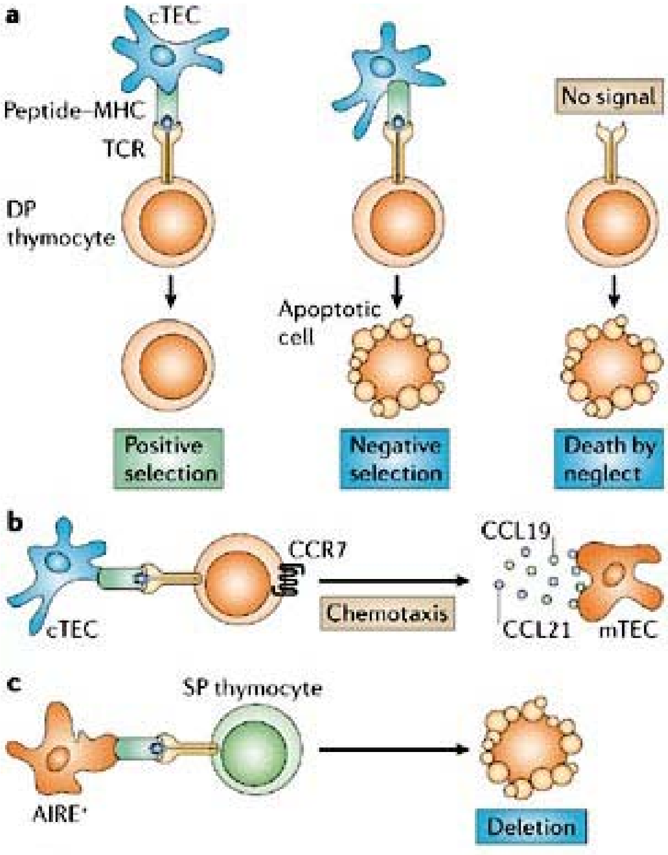


Figure 2 (a) Cortical thymic epithelial cells (cTECs) and dendritic cells present peptide-HLA complexes to double positive (DP) T lymphocytes in the cortex of the thymus. T lymphocytes are positively or negatively selected based on the interaction between their T cell receptor (TCR) and the peptide-HLA complex. Cells not receiving a survival signal die by apoptosis. (b) Positively selected DP T cells start to express CC-chemokine receptor 7 (CCR7), and migrated towards the medulla where medullar TECs (mTECs) express CC-chemokine ligand 19 (CCL19) and CC-chemokine ligand (CCL21), the ligands for CCR7. DP T cells undergo differentiation into single-positive lymphocytes. (c) mTECs express the transcriptional factor autoimmune regulator (AIRE) which promotes expression of tissue-specific antigens. mTECs present these antigens and self reactive T cells are deleted, hence central tolerance is maintained. Adapted from Takahama 2006.

Cathepsins

As described above, cysteine cathepsins are important players in the presentation of antigenic structure to immature T lymphocytes in the thymus, and hence in formation of a potent immune defence. Degradation of Ii and antigens are localized to the endosomal compartment of APCs, and cathepsins are responsible for this degradation (Figure 3). Cathepsins, in general, are proteases known to be involved in intracellular protein turnover, immune response and antigen processing, proprotein and hormone activation, remodelling of extracellular matrix and apoptosis (Berti and Storer 1995; Turk et al. 2000; Turk et al. 2001;

Turk et al. 2002; Friedrichs et al. 2003). The cathepsins found in the endosomal and lysosomal compartments of APCs are cysteine cathepsins. They are constitutively expressed as inactive preproenzymes, and there are 11 known cysteine cathepsins in humans (Turk et al. 2001; Rossi et al. 2004). The cathepsins have overlapping functions and different localizations. Protease inhibitors, like leupeptin, were initially used to identify which proteases were involved in Ii degradation. Cathepsin S was identified as the main cathepsin involved in Ii degradation outside the thymus (Riese et al. 1996; Nakagawa et al. 1999). In mice, cathepsin L was identified as the main cathepsin involved in Ii degradation in thymus (Nakagawa et al. 1998). Other cathepsins (B, C, H and X), are thought to trim amino- and carboxy-terminal ends of the peptides, either before or after the peptide binds the class II groove (Riese and Chapman 2000; Turk et al. 2001) (Figure 3).

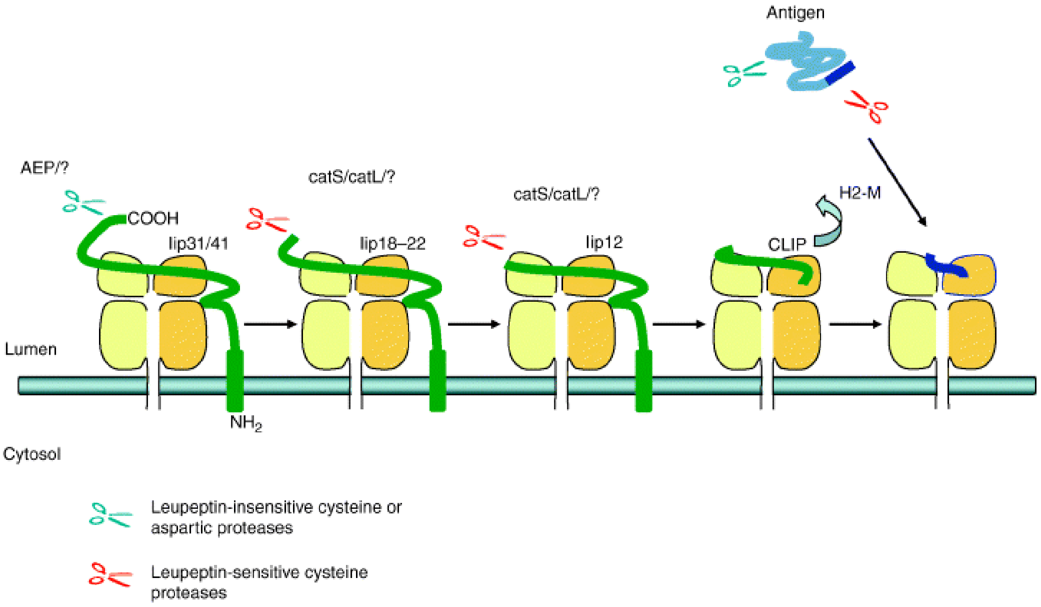

Antigen
AEP/?
catS/catL/?
catS/catL/?
H2-M
COOH
lip31/41
lip18–22
lip12
CLIP
Lumen
NH_2
Cytosol
Leupeptin-insensitive cysteine or aspartic proteases
Leupeptin-sensitive cysteine proteases

Figure 3 Degradation of invariant chain in late endosomes. Bound to the HLA class II heterodimer, the invariant chain (Ii) undergoes successive cleavage in the acidic environment of the endosomes of antigen presenting cells (APC). The initial cleavage is thought to be mediated by a leupeptin-insensitive cysteine protease or an aspartic protease, whereas subsequent steps are a result of the activity of cysteine cathepsins. For instance, in humans cathepsin S and cathepsin L2 are involved in the Ii degradation, mainly in APCs and cortical epithelial cells, respectively. Within the same endosomal compartment, cysteine and aspartic proteases also degrade internalized and endogenous proteins that, and the HLA-like molecule HLA-DM, exchanges class-II-associated invariant chain peptide (CLIP) for the peptides that are generated by these enzymes. Leupeptin is a protease inhibitor. Adapted from Hsing and Rudensky 2005.

CTSL2 as a candidate gene

Another important cysteine cathepsin is *CTSL2*; also called cathepsin V. This cathepsin is homologous to the murine cathepsin L. Murine cathepsin L has a unique role in the cortical part of the murine thymus, and it has been shown to be important for selection of CD4+ T cells (Nakagawa et al. 1998; Honey et al. 2002). In knockout mice lacking cathepsin L, the number of peripheral CD4+ T cells was dramatically reduced by 60-80% (Nakagawa et al. 1998).

CTSL2 was first reported to be expressed in testis and thymus, in addition to expression in breast and colorectal carcinomas (Santamaria et al. 1998). It was demonstrated that *CTSL2* belonged to the papain family of cysteine cathepsins, and that it showed 78% homology with human cathepsin L. The nucleotide sequence had the characteristic residues of the active site of cysteine cathepsins (Cys residue at amino acid position 138, His residue at position 277 and Asn at position 301). The 17 first amino acids were identified as a signal sequence, the next 96 amino acids as a pro-region, and the last 221 amino acids as the mature cathepsin (peptidase C1A domain). Peptidases hydrolyze peptide bonds, and the pro-region keeps the peptidase domain inactive until needed. In addition, the pro-region mediates proper protein folding and stabilization of the enzyme. Three potential N-glycosylation sites were identified, suggesting *CTSL2* to be targeted for lysosomes via the mannose-6-phosphate receptor system (Santamaria et al. 1998). Adachi and colleagues (1998) supported these findings, and showed expression of *CTSL2* in corneal epithelium. Brömme and colleagues (1999) confirmed the expression of *CTSL2* in thymus, but could not reproduce expression in a colorectal adenocarcinoma cell line. In addition, they reported no expression of *CTSL2* in other immune related organs like fetal liver, appendix, lymph nodes and bone marrow. *CTSL2* was mapped to chromosome 9q22.2, a site adjacent to cathepsin L. Because of the high homology between these two genes, a recent gene duplication from an ancestral cathepsin L-V like gene was suggested (Bromme et al. 1999). At that time, murine cathepsin L had been shown to be involved in degradation of Ii in thymic HLA class II processing (Nakagawa et al. 1998), and Brömme and colleagues (1999) speculate whether *CTSL2* could encounter the same role in human thymus, and hence be involved in positive T cell selection. Itoh and colleagues (1999) found that *CTSL2* consists of 8 exons and 7 introns, covering a length of approximately 6.6 kilo bases. They found the coding sequence to start within exon 2. The 5' untranslated region (UTR), which is 73 base pairs (bp) long, covers exon 1 and the first 10 bases of exon 2, while 3'UTR covers the 291 last bp of exon 8 (VEGA v16;

http://vega.sanger.ac.uk). Today, few alternative splice variants of *CTSL2* have been reported in the databases; in addition, none of the reported variants are from thymic tissue. Ensembl reports two known transcripts, one full length mRNA of 1457 bp with 8 exons, and one shorter variant, lacking exon 1, of 1294 bp with 7 exons (Ensembl v36; http://www.ensembl.org/index.html). VEGA also reports two different transcripts, one full length mRNA of 1369 bp with 8 exons, and one shorter variant consisting of exon 1-5 and only 577 bp long. NCBI only reports full length mRNA of 1496 bp (http://www.ncbi.nlm.nih.gov/).

As previously mentioned, *CTSL2* was shown to be expressed in cortical thymic epithelial cells (cTECs) (Tolosa et al. 2003), which are known to be involved in positive selection. *CTSL2* is now believed to be the main cathepsin involved in degradation of Ii in the thymus, analogous to cathepsin S outside the thymus (Zavasnik-Bergant and Turk 2006). Further, expression of *CTSL2* was suggested to be involved in the balance between positive and negative selection in the thymus (Tolosa et al. 2003). Interestingly, Tolosa and colleagues (2003) observed higher expression of *CTSL2* in thymi from MG patients with thymoma or thymitis when compared with controls (both healthy and patients with thymoma, but without MG). They speculated whether this could lead to a higher population of autoreactive T lymphocytes, and hence be involved in the autoimmune nature of the disease. Furthermore, the homolog murine cathepsin L has been suggested to play a role in development of diabetes in non-obese diabetic (NOD) mice (Maehr et al. 2005). They showed that Cathepsin L-deficient NOD mice were protected against diabetes, and that these mice displayed a defect in the generation of CD4+ T cells. Instead the proportion of regulatory T cells was increased, which suppressed the diabetogenic potential (Maehr et al. 2005). Interestingly, the chromosome area near the *CTSL2* gene in humans has previously been linked to T1D in a genome-wide screen for linkage in Scandinavia (Nerup and Pociot 2001). As described earlier, areas linked to one autoimmune disease might be of interest in other autoimmune diseases as well. Even though no such genome-wide screen for linkage has been performed for MG, *CTSL2* is nonetheless an interesting gene to further study in both T1D and MG. Especially to investigate whether the observed changes in expression can be related to polymorphisms regulating the gene and also the risk for disease development. T1D and MG have already been shown to share some common disease predisposing genes, and since *CTSL2* seems to be, in one way or another, involved in both diseases, it might turn out to be the next gene to be added on the list.

Gene expression and alternative splicing

Because of the high degree of LD (se below for explanation) often observed within confined genetic regions, unravelling whether an association is primary or secondary is challenging. In this regard, functional analyses of associated polymorphisms and candidate genes are often used to support interesting genetic findings. Gene expression studies, for instance, can be used to investigate whether an associated polymorphism could influence the transcription of the gene (Stranger et al. 2005). Areas of the genome earlier discarded as "junk DNA", like intronic sequences and intergenic regions, are now recognized as important holders of regulatory information. Only when we understand all the genetic control mechanisms present in a cell, the impact of polymorphisms is to be fully recognized.

Most studies investigate variants in *cis*-acting regions. These regions are often close to the gene, while *trans*-acting regions are further away, often at different chromosomes (Lettice et al. 2003; Pastinen et al. 2006). *Cis*-acting variants have been estimated to regulate 1-5% of all genes (Cheung et al. 2005; Pastinen et al. 2005; Stranger et al. 2005). These regulatory variants are often found within promoters and 5'UTRs, but also in coding and intronic regions. Accumulating evidence have pointed out 3'UTRs as just as important. The 3'UTRs of protein coding genes are often 3-5 times longer than the 5'UTRs, in addition they are rich in regulatory elements (reviwed in Chen et al. 2006). In 2005, 106 highly conserved motifs in 3'UTRs were identified. Few of these were known regulatory motifs, and the challenge will be to understand the specific functions of these (Xie et al. 2005). Studies identifying disease associated variants in 3'UTRs will lend insight into the regulatory aspects of these sequences.

Gene expression is under tight control in the cell, and there are many steps in the process from DNA to protein where correct regulation is crucial. The first point of regulation is transcriptional control; when and where a gene is transcribed. Further, the RNA needs to be correctly spliced and processed, in addition to correct transportation and localization in the cell. Protein translation and activity are also strictly regulated. Both *cis*- and *trans*-acting variants can affect all these steps, in addition to transcription factor binding, mRNA stability, nuclear export, translational efficiency and protein stability. Even small changes can have crucial impact on the organism. Predisposing polymorphisms can influence the fine tuning of the immune system, and accumulation of such risk variants might lead to autoimmune diseases in some individuals. For instance, SNPs can affect alternative splicing and give divergent products, which might influence disease susceptibility. Alternative splicing is part of the RNA processing that takes place in the nucleus. It is a way to create RNA and protein diversity, where the same transcript is alternatively spliced to produce several different transcripts and hence, potentially several different proteins. It has been estimated that between 40-74% of all human genes are subject to alternative splicing (Mironov et al. 1999; Brett et al. 2000; Kan et al. 2001; Modrek et al. 2001; Johnson et al. 2003). The different variants can be tissue, organ, developmental stage and disease specific. Tissue-specific splicing is most prevalent in brain cells (Stamm et al. 2000; Xu et al. 2002). The splicing machinery is tightly regulated. Genes are transcribed as pre-mRNA, and as much as 90% of the pre-mRNA sequence is removed by the splicing machinery (Stamm et al. 2005). Usually introns are spliced out during this RNA processing, leaving only the exons left to be translated. But alternative splicing can take many forms; e.g. whole or part of introns can be kept, exons can be spliced out or truncated, the length of 5' and 3' ends can vary. Alternative splicing can thus impact cell function in many ways.

For functional analyses, gene expression results can be correlated with the presence of associated polymorphisms in a population. Because disease predisposing polymorphisms usually are present in the general population, often at a high frequency, it is not necessary to study patients. Tissue from patients is often not easily collectable; in addition it can be an advantage to use healthy individuals since these are not undergoing immune-modulating treatment. When studying gene expression related to a disease, it is important to isolate RNA from the cells/tissue known, or believed, to express the gene and to be involved in disease pathogenesis.

Genetic polymorphisms used in mapping disease genes

Polymorphisms can be defined as any sites in the genome where multiple alleles exist as a stable component of the population. There are two different kinds of polymorphisms predominantly used in genetic studies today; SNPs and microsatellites, and both are abundant in the genome. Each SNP could potentially have four alleles (due to the four nucleotides), however most only have two. Microsatellites, on the other hand, have several different alleles.

Single nucleotide polymorphisms

SNPs are single nucleotide alterations in a genomic sequence (found in the general population), for instance a change from cytosine (C) to thymine (T), and where some individuals might have a C, others have a T. For most SNPs, the two allele variants present today is the result of a point mutation in distant past. Not all single nucleotide alterations are considered a SNP. A SNP can be defined as a single nucleotide exchange that is present in at least 1% of the population, however since allele frequencies vary between populations, this could also be a troublesome definition. SNPs are found all over the genome, in both coding and non-coding regions, and typically appear every few hundred bp (Kwok et al. 1996; Nickerson et al. 2000; Altshuler et al. 2005). There are more than 10 million SNPs deposited in the human dbSNP database so far, and 5.8 millions of these are validated (dbSNP; http://www.ncbi.nlm.nih.gov/projects/SNP/). Most of the identified SNPs are in non-coding regions, and earlier these SNPs were believed to be of no interest. The primary goal of genetic studies was to identify SNPs in coding regions that would lead to amino acid substitution, and subsequently have an impact on diseases pathogenesis. Several such polymorphisms were identified in monogenic diseases like cystic fibrosis (CF) and hereditary haemocromatosis (HH). In CF, one deleterious polymorphism was found in 70% of the patients, leading to a loss of phenylalanine in the cystic fibrosis transmembrane conductance regulator (CFTR) (Kerem et al. 1989). The consequence is defective intracellular transport and incomplete processing of CFTR (Cheng et al. 1990). The last 30% of the CF patients have several different known polymorphisms in the *CFTR* gene. In HH on the other hand, one major polymorphism in the haemocromatosis (*HFE*) gene leading to an amino acid substitution, and further inactivation of the protein, is associated with the diseases (Feder et al. 1996).

When studying complex autoimmune diseases, even small effects, like changes in gene expression or mRNA stability due to a SNP in a regulatory sequence, are of high interest. As previously mentioned, SNPs outside the coding regions have been shown to be involved in disease predisposition, like the SNP in the gene encoding *CTLA4* that is associated with T1D (Ueda et al. 2003; Bottini et al. 2004). When selecting SNPs for genetic studies, both SNPs in coding regions and non-coding regions should be considered. In order to increase the chances of finding functional SNPs in non-coding sequences, the gene area under investigation should be compared in several different species. Recent comparisons of mammalian genomes have revealed highly conserved regions that contain no coding sequences (reviewed in Dermitzakis et al. 2005). These sequences are called conserved non-genic sequences (CNG) or conserved non-coding sequences (CNS) (Meisler 2001; Dermitzakis et al. 2005). Regions conserved in several species are more likely to contain important regulatory sequences (Loots et al. 2000), however it seems that these regions encompass more than cis-acting transcriptional regulators (reviewed in Dermitzakis et al. 2005). Delineating all the functions of intergenic and intronic sequences is one of the challenges in the future, and today, SNPs in non-coding regions are studied on a broad scale.

Microsatellites

As mentioned, microsatellites are also widely used genetic polymorphisms in mapping studies. Microsatellites consist of a unit of nucleotides (2-4 bp) that is repeated a variable number of times. Alleles are usually distinguished and named after the length of the polymerase chain reaction (PCR) product (ex: 119, 121, 123 bp etc). Microsatellites are highly polymorphic, and like SNPs, abundant all over the genome (Dib et al. 1996). Few microsatellites have been shown to have functional implications, but because of the high degree of heterozygosity, they are very informative for the mapping process. Where SNPs usually have two alleles, microsatellites have several alleles, making them good tools to pick up associated haplotypes and SNPs through LD. Because they consist of numerous alleles, they could potentially distinguish associated from non-associated haplotypes.

Factors affecting genetic polymorphisms

Identifying genetic polymorphisms that might influence disease development, as described above, are challenging. Genetic studies try to identify differences in allele frequencies in two groups (patients and controls) that are not due to normal distribution of allele frequencies. Knowledge of the parameters influencing allele frequencies in a population is important, and some of the most essential are;

- Population size; a large population displays a wide variety of polymorphisms, while a smaller population generally displays less variety. If a population is exposed to a severe pathogen, individuals with polymorphisms protecting against the infection are more likely to survive, and a selective pressure is exercised on that polymorphism (also called selection). Another example is bottlenecks; several populations have experienced bottlenecks e.g. due to famine, slavery, infections. As a consequence inbreeding, genetic drift and possible selection might change the allelic distributions.
- Random mating; when a population mates randomly, a great variety of polymorphisms are kept. Inbreeding leads to accumulation of fewer variants of a given polymorphism.
- Migration and admixture; when two populations with different allele frequencies are mixed, a shift will be detectable when frequencies are compared before and after the mixing.
- Mutation; for a random mutation to be kept in a population, it must either not lead to any changes or be an advantage for the individual, and it must reach fixation.

These factors influence allele frequencies and LD between alleles, and thereby our search for disease predisposing genes.

Linkage disequilibrium

When no disease affecting polymorphism is known prior to a study, researchers rely on the presence of LD between the polymorphisms tested and the undetermined disease associated allele. In general, when alleles at two different loci are in linkage equilibrium, they are randomly combined on chromosomes in the population. This occurs if two loci are either located on two different chromosomes or sufficiently separated on the same chromosome so that recombination has had time to drive the haplotypes to random frequencies. Recombination frequency describes the rate of crossing over of genetic material between chromosomes during meiosis. As previously mentioned, when considering a SNP present today with two different alleles; at one point in history a single point mutation occurred and a new variant arose. The new SNP arose on a particular haplotype background. Over time, the gene region comprising this haplotype might have undergone recombination, gene conversion or even new mutations, and at present time, several different haplotypes including one or the other allele might exist (Figure 4) (reviewed in Abecasis et al. 2005). Other SNPs at nearby loci might have undergone the same processes.

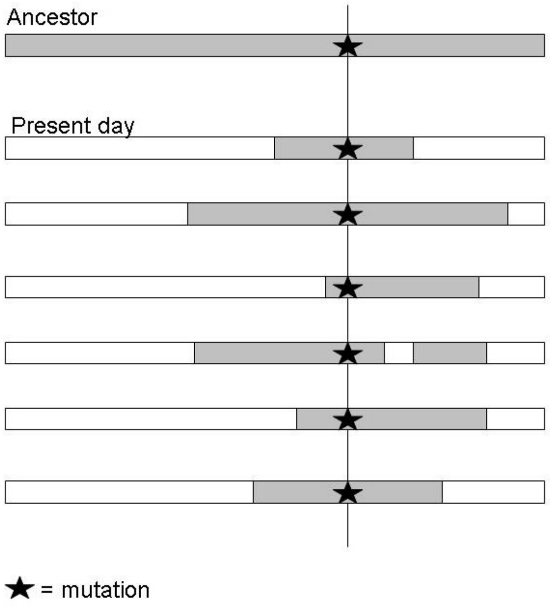


Figure 4 Schematic presentation of linkage disequilibrium. Most SNPs originate as point mutations. Here presented on a hypothetical ancestral chromosome colored grey. Present day individuals carrying the mutant allele will also carry a surrounding stretch of the ancestral chromosome of variable size. Alleles within these short stretches are not associated randomly, but instead often appear in the same configuration as in the original mutant chromosome, and are said to be in linkage disequilibrium. The exact length of this conserved stretch will depend on local rates of recombination, gene conversion, and mutation. Population history, especially through genetic drift and natural selection, also plays a role as it shapes the geneaology of individual alleles and determines the number of rounds of mutation, recombination and natural selection to which each allele is subject. Adapted from Abecasis et al. 2005.

When two alleles on the same chromosome are close together, they are less likely to be dissolved during meiosis, and can be transmitted together as a haplotype. When alleles at two loci appear together as a haplotype more often than expected, the alleles are in LD. D' and r^2 are often used as measures for the degree of LD between two loci (reviewed in Zondervan and Cardon 2004). Both describe the difference between the probability of observing two alleles on the same haplotype and observing them independently in the population (Figure 5). Two alleles always present on the same haplotype have a D' value of 1 (complete LD). When two alleles are never seen on the same haplotype, the D' value is -1, and when two alleles are in linkage equilibrium the D' value is 0. An r^2 value of 0 also implies independence, however r^2=1 is more strictly defined; only when the allele loci have identical allele frequencies and every occurrence of an allele at each of the loci perfectly predicts the allele at the other locus (Figure 5).

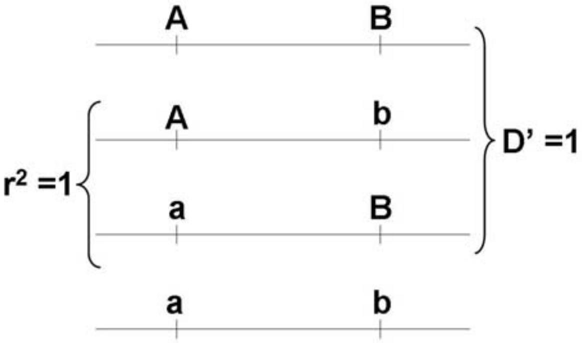


Figure 5 Schematic illustration of the difference between the LD parameters D' and r^2. Two SNPs (A/a and B/b) can potentially give rise to four possible haplotypes in a population (indicated by horizontal lines). The LD measurement r^2 is bidirectional. Hence, when r^2=1 only two of the four haplotypes are present in the population and the two SNPs are fully correlated (meaning: if A is present, b is always carried on the same haplotype and the other way around). As a consequence, when r^2=1 the allele frequencies of the two polymorphisms are equal and the same as their joint haplotype frequency. D' is a unidirectional test and less stringent. When D'=1 three of the four possible haplotypes are present in the population. Three random selected haplotypes are shown in the figure, and in this situation only if an individual carries either a or b, can the allele at the other position be predicted with 100% certainty, while this prediction is not possible for individuals carrying B or A.

The fact that polymorphisms are in LD with each other can be used as a tool to reduce the number of polymorphisms needed to genetically map a region of the genome. Two SNPs in complete LD will not be more informative than each SNP by itself.

Statistical considerations

When performing genetic studies, the goal is to identify genetic variants that might tell us why some individuals are predisposed for a specific disease. Genetic studies are widely discussed, and the main reason is because many positive studies have failed to be reproduced.

This applies to both genome-wide and candidate gene studies. Several issues make these studies difficult to perform. In complex autoimmune diseases the symptoms often vary, the age of onset varies and the presence of biological markers (like antibodies) might differ, hence defining groups of patients based on phenotype might be troublesome. The molecular pathway(s) involved in the diseases might be different within a group of patients; this is again due to the heterogeneity often seen in complex diseases. In addition, several genes often contribute to disease susceptibility, each with small individual contributions and relative risks (Tabor et al. 2002).

Statistical methods applied in association studies

To identify genetic association, allele frequencies of a given polymorphism are compared between two groups (patients and controls), and significant differences might indicate association with the disease, either directly or indirectly through LD. Typically, differences between patients and controls that give rise to a p-values<0.05, are considered significant. However, to avoid false positive associations, p-values should be corrected for multiple testing (number of polymorphisms tested and number of subgroup analysis performed) (Colhoun et al. 2003) and/or the initially reported association should be replicated in an independent data set.

Mendelian inheritance of the polymorphisms by utilizing the PedCheck program is often applied on family materials to confirm that the individuals are related, and to detect marker typing incompatibilities within the family. In addition, the genetic polymorphisms used should be tested to see whether they are in Hardy-Weinberg equilibrium (HWE) or not. When polymorphisms are in HWE, the observed genotype frequencies are as expected from their allele frequencies. HWE is used as a quality control of the genotyping (Hosking et al. 2004), as well as a requirement for statistical tests of allelic associations.

When planning a genetic association study, both case-control and family materials can be used. A challenge with case-control studies is how well the controls ethnically match the cases. For instance, if the population is ethnically mixed, differences in allele frequencies might arise because the cases and controls consist of different proportions of the two original populations. This is called population stratification. To avoid this stratification problem, family-based materials can be used, where a genetically matched and family-based control for each case is used. The drawback of family materials compared with case-control materials, is the fact that twice as many “controls” need to be genotyped (two parental “controls” pr case vs. one unrelated control pr case).

A widely used test in family materials is the transmission disequilibrium test (TDT), a test that determines if a particular allele is associated with disease (Spielman et al. 1993). TDT considers transmission of alleles from heterozygous parents to affected offspring (proband). Assuming no association, there is a 50% chance of each allele to be transmitted, and deviations from the expected transmission of 50% are calculated using McNemar statistics (Spielman et al. 1993). TDT is biallelic, i.e. if there are more that two alleles; each allele is tested against all others. Another widely used test in family studies is the affected family-based controls (AFBAC) test (Thomson 1995). Here, transmissions from all parents are included, not only from the heterozygous. The transmitted alleles represent the case population, while the non-transmitted alleles represent the control population. It has been shown that the latter frequency resembles that seen in the general population. Because the transmissions are not treated as a pair in this test, population stratification is not avoided.

For a case-control material, the frequencies observed in cases and controls are compared using chi-squared (chi^2) statistics to determine whether the differences are significant. This test calculates the deviation between the expected (no association) and the observed allele frequencies.

Power and odds ratio

Whether one is able to detect significant association depends on the effect of the risk allele, the allele frequencies in the population studied, and how many cases and controls, or families, the study includes. This is defined as power. Power is restricted by the smallest group; hence at least as many controls as patients are preferred. If results are negative, power calculations becomes of high importance. Are the results truly negative, i.e. no association, or is it due to lack of power and thereby a false negative finding? When interpreting association data, the pitfalls of both false negative and false positive results should be considered. When genotyping for instance 100 SNPs, by chance five of these will come out as false positive. Therefore, either correction for multiple tests or confirmation of the results is warranted.

As mentioned earlier, complex diseases are often a result of polymorphisms in several genes and each polymorphism usually contributes only with a small risk. Odds ratio (OR) is a measure of disease risk. This method describes the ratio between two odds, namely the odds of disease if carrying the presumed disease associated allele over the other odds of disease if not carrying the disease associated allele. An OR=1 describes no relationship between disease and the allele studied, while OR>1 means increased risk of diseases and OR<1 means protection against the disease. OR-values should be given with a 95% confidence interval, and significantly distorted risk is observed if the confidence interval does not cover 1.

As mentioned previously, autoimmune diseases are caused by the interaction of several genetic loci. The risk most genes hold are small, making the identification of susceptibility genes difficult. The three non-MHC genes linked to T1D have low OR values; *INS* (OR= ~1.9), *CTLA4* (OR= ~1.2) and *PTPN22* (OR= ~1.7) (Concannon et al. 2005). Therefore, the importance of power becomes evident, as most susceptibility genes seem to have relatively low OR values.

Aim of the present study

Complex autoimmune diseases often share some common disease susceptibility genes, like the identified *CTLA4* and *PTPN22* genes, which are found to be associated with T1D, MG, SLE and RA. *CTSL2* has previously been shown to be up regulated in thymic tissue from MG patients. In addition, the gene has been suggested to play a role in development of diabetes in non-obese diabetic (NOD) mice. *CTSL2* is known to be involved in antigen processing and positive selection in thymus, hence an interesting gene for autoimmunity in general.

This thesis aims at delineating the following questions:

1. Is *CTSL2* associated with T1D and/or MG?
2. Do alternative transcripts of *CTSL2* exist in thymic tissue?
3. Is there any correlation between disease associated polymorphisms (if identified) and gene expression of *CTSL2* in thymic tissue?

References

Abecasis GR, Ghosh D, Nichols TE (2005) Linkage disequilibrium: ancient history drives the new genetics. Hum Hered 59:118-124

Adachi W, Kawamoto S, Ohno I, Nishida K, Kinoshita S, Matsubara K, Okubo K (1998) Isolation and characterization of human cathepsin V: a major proteinase in corneal epithelium. Invest Ophthalmol Vis Sci 39:1789-1796

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299-1320

Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, Turley SJ, von Boehmer H, Bronson R, Dierich A, Benoist C, Mathis D (2002) Projection of an immunological self shadow within the thymus by the aire protein. Science 298:1395-1401

Aspinall R, Andrew D (2000) Thymic involution in aging. J Clin Immunol 20:250-256

Barnett AH, Eff C, Leslie RD, Pyke DA (1981) Diabetes in identical twins. A study of 200 pairs. Diabetologia 20:87-93

Becker KG (1999) Comparative genetics of type 1 diabetes and autoimmune disease: common loci, common pathways? Diabetes 48:1353-1358

Becker KG, Simon RM, Bailey-Wilson JE, Freidlin B, Biddison WE, McFarland HF, Trent JM (1998) Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. Proc Natl Acad Sci U S A 95:9979-9984

Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, et al. (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am J Hum Genet 75:330-337

Bennett ST, Lucassen AM, Gough SC, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, et al. (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9:284-292

Berti PJ, Storer AC (1995) Alignment/phylogeny of the papain superfamily of cysteine proteases. J Mol Biol 246:273-283

Bias WB, Reveille JD, Beaty TH, Meyers DA, Arnett FC (1986) Evidence that autoimmunity in man is a Mendelian dominant trait. Am J Hum Genet 39:584-602

Bottini N, Musumeci L, Alonso A, Rahmouni S, Nika K, Rostamkhani M, MacMurray J, Meloni GF, Lucarelli P, Pellecchia M, Eisenbarth GS, Comings D, Mustelin T (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat Genet 36:337-338

Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett 474:83-86

Bromme D, Li Z, Barnes M, Mehler E (1999) Human cathepsin V functional expression, tissue distribution, electrostatic surface potential, enzymatic characterization, and chromosomal localization. Biochemistry 38:2377-2385

Caillat-Zucman S, Garchon HJ, Timsit J, Assan R, Boitard C, Djilali-Saiah I, Bougneres P, Bach JF (1992) Age-dependent HLA genetic heterogeneity of type 1 insulin-dependent diabetes mellitus. J Clin Invest 90:2242-2250

Candore G, Lio D, Colonna Romano G, Caruso C (2002) Pathogenesis of autoimmune diseases associated with 8.1 ancestral haplotype: effect of multiple gene interactions. Autoimmun Rev 1:29-35

Carlsson B, Wallin J, Pirskanen R, Matell G, Smith CI (1990) Different HLA DR-DQ associations in subgroups of idiopathic myasthenia gravis. Immunogenetics 31:285-290

Chen JM, Ferec C, Cooper DN (2006) A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview. Hum Genet. In press.

Cheng SH, Gregory RJ, Marshall J, Paul S, Souza DW, White GA, O'Riordan CR, Smith AE (1990) Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. Cell 63:827-834

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365-1369

Cohen S, Dadi H, Shaoul E, Sharfe N, Roifman CM (1999) Cloning and characterization of a lymphoid-specific, inducible human protein tyrosine phosphatase, Lyp. Blood 93:2013-2024

Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361:865-872

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci U S A 96:15173-15177

Compston DA, Vincent A, Newsom-Davis J, Batchelor JR (1980) Clinical, pathological, HLA antigen and immunological evidence for disease heterogeneity in myasthenia gravis. Brain 103:579-601

Concannon P, Erlich HA, Julier C, Morahan G, Nerup J, Pociot F, Todd JA, Rich SS (2005) Type 1 diabetes: evidence for susceptibility loci from four genome-wide linkage scans in 1,435 multiplex families. Diabetes 54:2995-3001

Correa PA, Gomez LM, Cadena J, Anaya JM (2005) Autoimmunity and tuberculosis. Opposite association with TNF polymorphism. J Rheumatol 32:219-224

Cresswell P (1998) Proteases, processing, and thymic selection. Science 280:394-395

Cucca F, Lampis R, Frau F, Macis D, Angius E, Masile P, Chessa M, Frongia P, Silvetti M, Cao A, et al. (1995) The distribution of DR4 haplotypes in Sardinia suggests a primary association of type I diabetes with DRB1 and DQB1 loci. Hum Immunol 43:301-308

Dawkins RL, Christiansen FT, Kay PH, Garlepp M, McCluskey J, Hollingsworth PN, Zilko PJ (1983) Disease associations with complotypes, supratypes and haplotypes. Immunol Rev 70:1-22

Derbinski J, Gabler J, Brors B, Tierling S, Jonnakuty S, Hergenhahn M, Peltonen L, Walter J, Kyewski B (2005) Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. J Exp Med 202:33-45

Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet 6:151-157

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152-154

Drachman DB (1994) Myasthenia gravis. N Engl J Med 330:1797-1810

Egerton M, Scollay R, Shortman K (1990) Kinetics of mature T-cell development in the thymus. Proc Natl Acad Sci U S A 87:2579-2582

Engel AG (1984) Myasthenia gravis and myasthenic syndromes. Ann Neurol 16:519-534

Engel AG, Ohno K, Sine SM (2003) Congenital myasthenic syndromes: progress over the past decade. Muscle Nerve 27:4-25

Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, et al. (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399-408

Feltkamp TE, van den Berg-Loonen PM, Nijenhuis LE, Engelfriet CP, van Rossum AL, van Loghem JJ, Oosterhuis HJ (1974) Myasthenia gravis, autoantibodies, and HL-A antigens. Br Med J 1:131-133

Friedrichs B, Tepel C, Reinheckel T, Deussing J, von Figura K, Herzog V, Peters C, Saftig P, Brix K (2003) Thyroid functions of mouse cathepsins B, K, and L. J Clin Invest 111:1733-1745

Gallegos AM, Bevan MJ (2004) Central tolerance to tissue-specific antigens mediated by direct and indirect antigen presentation. J Exp Med 200:1039-1049

Garchon HJ, Djabiri F, Viard JP, Gajdos P, Bach JF (1994) Involvement of human muscle acetylcholine receptor alpha-subunit gene (CHRNA) in susceptibility to myasthenia gravis. Proc Natl Acad Sci U S A 91:4668-4672

Giraud M, Beaurain G, Yamamoto AM, Eymard B, Tranchant C, Gajdos P, Garchon HJ (2001) Linkage of HLA to myasthenia gravis and genetic heterogeneity depending on anti-titin antibodies. Neurology 57:1555-1560

Goldrath AW, Bevan MJ (1999) Selecting and maintaining a diverse T-cell repertoire. Nature 402:255-262

Green A, Patterson CC (2001) Trends in the incidence of childhood-onset diabetes in Europe 1989-1998. Diabetologia 44 Suppl 3:B3-8

Haynes BF, Heinly CS (1995) Early human T cell development: analysis of the human thymus at the time of initial entry of hematopoietic stem cells into the fetal thymic microenvironment. J Exp Med 181:1445-1458

Hill RJ, Zozulya S, Lu YL, Ward K, Gishizky M, Jallal B (2002) The lymphoid protein tyrosine phosphatase Lyp interacts with the adaptor molecule Grb2 and functions as a negative regulator of T-cell activation. Exp Hematol 30:237-244

Hjelmstrom P, Peacock CS, Giscombe R, Pirskanen R, Lefvert AK, Blackwell JM, Sanjeevi CB (1998) Polymorphism in tumor necrosis factor genes associated with myasthenia gravis. J Neuroimmunol 88:137-143

Hoch W, McConville J, Helms S, Newsom-Davis J, Melms A, Vincent A (2001) Auto-antibodies to the receptor tyrosine kinase MuSK in patients with myasthenia gravis without acetylcholine receptor antibodies. Nat Med 7:365-368

Hohlfeld R, Wekerle H (1994) The role of the thymus in myasthenia gravis. Adv Neuroimmunol 4:373-386

Honey K, Benlagha K, Beers C, Forbush K, Teyton L, Kleijmeer MJ, Rudensky AY, Bendelac A (2002) Thymocyte expression of cathepsin L is essential for NKT cell development. Nat Immunol 3:1069-1074

Honey K, Rudensky AY (2003) Lysosomal cysteine proteases regulate antigen presentation. Nat Rev Immunol 3:472-482

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC, Jr., Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889-899

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet 12:395-399

Hsing LC, Rudensky AY (2005) The lysosomal cysteine proteases in MHC class II antigen presentation. Immunological Reviews 207:229-241

Huang D, Liu L, Noren K, Xia SQ, Trifunovic J, Pirskanen R, Lefvert AK (1998) Genetic association of Ctla-4 to myasthenia gravis with thymoma. J Neuroimmunol 88:192-198

Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG (2003) Genetic associations in large versus small studies: an empirical assessment. Lancet 361:567-571

Itoh R, Kawamoto S, Adachi W, Kinoshita S, Okubo K (1999) Genomic organization and chromosomal localization of the human cathepsin L2 gene. DNA Res 6:137-140

Jacobson DL, Gange SJ, Rose NR, Graham NM (1997) Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol 84:223-243

Jameson SC, Hogquist KA, Bevan MJ (1995) Positive selection of thymocytes. Annu Rev Immunol 13:93-126

Johansson S, Lie BA, Todd JA, Pociot F, Nerup J, Cambon-Thomsen A, Kockum I, Akselsen HE, Thorsby E, Undlien DE (2003) Evidence of at least two type 1 diabetes susceptibility genes in the HLA complex distinct from HLA-DQB1, -DQA1 and -DRB1. Genes Immun 4:46-53

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141-2144

Kan Z, Rouchka EC, Gish WR, States DJ (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res 11:889-900

Karvonen M, Pitkaniemi J, Tuomilehto J (1999) The onset age of type 1 diabetes in Finnish children has become younger. The Finnish Childhood Diabetes Registry Group. Diabetes Care 22:1066-1070

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-1080

Kisielow P, Teh HS, Bluthmann H, von Boehmer H (1988) Positive selection of antigen-specific T cells in thymus by restricting MHC molecules. Nature 335:730-733

Kwok PY, Deng Q, Zakeri H, Taylor SL, Nickerson DA (1996) Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. Genomics 31:123-126

Kyogoku C, Langefeld CD, Ortmann WA, Lee A, Selby S, Carlton VE, Chang M, Ramos P, Baechler EC, Batliwalla FM, Novitzke J, Williams AH, Gillett C, Rodine P, Graham RR, Ardlie KG, Gaffney PM, Moser KL, Petri M, Begovich AB, Gregersen PK, Behrens TW (2004) Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. Am J Hum Genet 75:504-507

Kyvik KO, Green A, Beck-Nielsen H (1995) Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins. Bmj 311:913-917

Labeit S, Kolmerer B (1995) Titins: giant proteins in charge of muscle ultrastructure and elasticity. Science 270:293-296

Ladner MB, Bottini N, Valdes AM, Noble JA (2005) Association of the single nucleotide polymorphism C1858T of the PTPN22 gene with type 1 diabetes. Hum Immunol 66:60-64

Lander ES (1996) The new genomics: global views of biology. Science 274:536-539

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037-2048

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12:1725-1735

Lie BA, Thorsby E (2005) Several genes in the extended human MHC contribute to predisposition to autoimmune diseases. Curr Opin Immunol 17:526-531

Lie BA, Todd JA, Pociot F, Nerup J, Akselsen HE, Joner G, Dahl-Jorgensen K, Ronningen KS, Thorsby E, Undlien DE (1999) The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non-class II gene. Am J Hum Genet 64:793-800

Lindstrom JM, Seybold ME, Lennon VA, Whittingham S, Duane DD (1976) Antibody to acetylcholine receptor in myasthenia gravis. Prevalence, clinical correlates, and diagnostic value. Neurology 26:1054-1059

Liston A, Lesage S, Wilson J, Peltonen L, Goodnow CC (2003) Aire regulates negative selection of organ-specific T cells. Nat Immunol 4:350-354

Liu MF, Liu HS, Wang CR, Lei HY (1998) Expression of CTLA-4 molecule in peripheral blood T lymphocytes from patients with systemic lupus erythematosus. J Clin Immunol 18:392-398

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177-182

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136-140

Maehr R, Mintern JD, Herman AE, Lennon-Dumenil AM, Mathis D, Benoist C, Ploegh HL (2005) Cathepsin L is essential for onset of autoimmune diabetes in NOD mice. J Clin Invest 115:2934-2943

Meisler MH (2001) Evolutionarily conserved noncoding DNA in the human genome: how much and what for? Genome Res 11:1617-1618

Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. Genome Res 9:1288-1293

Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29:2850-2859

Mossman S, Vincent A, Newsom-Davis J (1986) Myasthenia gravis without acetylcholine-receptor antibody: a distinct disease entity. Lancet 1:116-119

Nagamine K, Peterson P, Scott HS, Kudoh J, Minoshima S, Heino M, Krohn KJ, Lalioti MD, Mullis PE, Antonarakis SE, Kawasaki K, Asakawa S, Ito F, Shimizu N (1997) Positional cloning of the APECED gene. Nat Genet 17:393-398

Nakagawa T, Roth W, Wong P, Nelson A, Farr A, Deussing J, Villadangos JA, Ploegh H, Peters C, Rudensky AY (1998) Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus. Science 280:450-453

Nakagawa TY, Brissette WH, Lira PD, Griffiths RJ, Petrushova N, Stock J, McNeish JD, Eastman SE, Howard ED, Clarke SR, Rosloniec EF, Elliott EA, Rudensky AY (1999) Impaired invariant chain degradation and antigen presentation and diminished collagen-induced arthritis in cathepsin S null mice. Immunity 10:207-217

Nerup J, Pociot F (2001) A genomewide scan for type 1-diabetes susceptibility in Scandinavian families: identification of new loci with evidence of interactions. Am J Hum Genet 69:1301-1313

Nickerson DA, Taylor SL, Fullerton SM, Weiss KM, Clark AG, Stengard JH, Salomaa V, Boerwinkle E, Sing CF (2000) Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Res 10:1532-1545

Onengut-Gumuscu S, Ewens KG, Spielman RS, Concannon P (2004) A functional polymorphism (1858C/T) in the PTPN22 gene is linked and associated with type I diabetes in multiplex families. Genes Immun 5:678-680

Oosterhuis H (1997) Myasthenia Gravis. Groningen: Groningen Neurological Press

Owen JJ, Ritter MA (1969) Tissue interaction in the development of thymus lymphocytes. J Exp Med 129:431-442

Owerbach D, Lernmark A, Platz P, Ryder LP, Rask L, Peterson PA, Ludvigsson J (1983) HLA-D region beta-chain DNA endonuclease fragments differ between HLA-DR identical healthy and insulin-dependent diabetic individuals. Nature 303:815-817

Pastinen T, Ge B, Gurd S, Gaudin T, Dore C, Lemire M, Lepage P, Harmsen E, Hudson TJ (2005) Mapping common regulatory variants to human haplotypes. Hum Mol Genet 14:3963-3971

Pastinen T, Ge B, Hudson TJ (2006) Influence of human genome polymorphism on gene expression. Hum Mol Genet 15:R9-R16

Patrick J, Lindstrom J (1973) Autoimmune response to acetylcholine receptor. Science 180:871-872

Pirskanen R (1976) Genetic associations between myasthenia gravis and the HL-A system. J Neurol Neurosurg Psychiatry 39:23-33

Price P, Witt C, Allcock R, Sayer D, Garlepp M, Kok CC, French M, Mallal S, Christiansen F (1999) The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol Rev 167:257-274

Pugliese A, Gianani R, Moromisato R, Awdeh ZL, Alper CA, Erlich HA, Jackson RA, Eisenbarth GS (1995) HLA-DQB1*0602 is associated with dominant protection from diabetes even among islet cell antibody-positive first-degree relatives of patients with IDDM. Diabetes 44:608-613

Pugliese A, Zeller M, Fernandez A, Jr., Zalcberg LJ, Bartlett RJ, Ricordi C, Pietropaolo M, Eisenbarth GS, Bennett ST, Patel DD (1997) The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. Nat Genet 15:293-297

Qu H, Tessier MC, Hudson TJ, Polychronakos C (2005) Confirmation of the association of the R620W polymorphism in the protein tyrosine phosphatase PTPN22 with type 1 diabetes in a family based study. J Med Genet 42:266-270

Redondo MJ, Fain PR, Eisenbarth GS (2001) Genetics of type 1A diabetes. Recent Prog Horm Res 56:69-89

Riese RJ, Chapman HA (2000) Cathepsins and compartmentalization in antigen presentation. Curr Opin Immunol 12:107-113

Riese RJ, Wolf PR, Bromme D, Natkin LR, Villadangos JA, Ploegh HL, Chapman HA (1996) Essential role for cathepsin S in MHC class II-associated invariant chain processing and peptide loading. Immunity 4:357-366

Risch N (1990a) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228

Risch N (1990b) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229-241

Risch N (1990c) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242-253

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

Romi F, Gilhus NE, Aarli JA (2005) Myasthenia gravis: clinical, immunological, and therapeutic advances. Acta Neurol Scand 111:134-141

Rossi A, Deveraux Q, Turk B, Sali A (2004) Comprehensive search for cysteine cathepsins in the human genome. Biol Chem 385:363-372

Santamaria I, Velasco G, Cazorla M, Fueyo A, Campo E, Lopez-Otin C (1998) Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. Cancer Res 58:1624-1630

Scollay RG, Butcher EC, Weissman IL (1980) Thymus cell migration. Quantitative aspects of cellular traffic from the thymus to the periphery in mice. Eur J Immunol 10:210-218

Seidl C, Donner H, Fischer B, Usadel KH, Seifried E, Kaltwasser JP, Badenhoop K (1998) CTLA4 codon 17 dimorphism in patients with rheumatoid arthritis. Tissue Antigens 51:62-66

Sheehy MJ, Scharf SJ, Rowe JR, Neme de Gimenez MH, Meske LM, Erlich HA, Nepom BS (1989) A diabetes-susceptible HLA haplotype is best defined by a combination of HLA-DR and -DQ alleles. J Clin Invest 83:830-835

Smyth D, Cooper JD, Collins JE, Heward JM, Franklyn JA, Howson JM, Vella A, Nutland S, Rance HE, Maier L, Barratt BJ, Guja C, Ionescu-Tirgoviste C, Savage DA, Dunger DB, Widmer B, Strachan DP, Ring SM, Walker N, Clayton DG, Twells RC, Gough SC, Todd JA (2004) Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. Diabetes 53:3020-3023

Spielman RS, Baker L, Zmijewski CM (1980) Gene dosage and suceptibility to insulin-dependent diabetes. Ann Hum Genet 44:135-150

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

Spurkland A, Gilhus NE, Ronningen KS, Aarli JA, Vartdal F (1991) Myasthenia gravis patients with thymus hyperplasia and myasthenia gravis patients with thymoma display different HLA associations. Tissue Antigens 37:90-93

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H (2005) Function of alternative splicing. Gene 344:1-20

Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ (2000) An alternative-exon database and its statistical analysis. DNA Cell Biol 19:739-756

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-Wide Associations of Gene Expression Variation in Humans. PLoS Genet 1:e78

Su MA, Anderson MS (2004) Aire: an update. Curr Opin Immunol 16:746-752

Suarez BK (1978) The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. Tissue Antigens 12:87-93

Tabor HK, Risch NJ, Myers RM (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3:391-397

Takahama Y (2006) Journey through the thymus: stromal guides for T-cell development and selection. Nat Rev Immunol 6:127-135

The MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. Nature 401:921-923

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487-498

Thorlacius S, Aarli JA, Riise T, Matre R, Johnsen HJ (1989) Associated disorders in myasthenia gravis: autoimmune diseases and their relation to thymectomy. Acta Neurol Scand 80:290-295

Thorsby E, Lie BA (2005) HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. Transpl Immunol 14:175-182

Tisch R, McDevitt H (1996) Insulin-Dependent Diabetes Mellitus Cell 85:291-297

Tolosa E, Li W, Yasuda Y, Wienhold W, Denzin LK, Lautwein A, Driessen C, Schnorrer P, Weber E, Stevanovic S, Kurek R, Melms A, Bromme D (2003) Cathepsin V is involved in the degradation of invariant chain in human thymus and is overexpressed in myasthenia gravis. J Clin Invest 112:517-526

Turk B, Stoka V, Rozman-Pungercar J, Cirman T, Droga-Mazovec G, Oresic K, Turk V (2002) Apoptotic pathways: involvement of lysosomal proteases. Biol Chem 383:1035-1044

Turk B, Turk D, Turk V (2000) Lysosomal cysteine proteases: more than scavengers. Biochim Biophys Acta 1477:98-111

Turk V, Turk B, Turk D (2001) Lysosomal cysteine proteases: facts and opportunities. Embo J 20:4629-4633

Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, et al. (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 423:506-511

Ueno T, Saito F, Gray DH, Kuse S, Hieshima K, Nakano H, Kakiuchi T, Lipp M, Boyd RL, Takahama Y (2004) CCR7 signals are essential for cortex-medulla migration of developing thymocytes. J Exp Med 200:493-505

Undlien DE, Bennett ST, Todd JA, Akselsen HE, Ikaheimo I, Reijonen H, Knip M, Thorsby E, Ronningen KS (1995) Insulin gene region-encoded susceptibility to IDDM maps upstream of the insulin gene. Diabetes 44:620-625

Undlien DE, Friede T, Rammensee HG, Joner G, Dahl-Jorgensen K, Sovik O, Akselsen HE, Knutsen I, Ronningen KS, Thorsby E (1997) HLA-encoded genetic predisposition in IDDM: DR4 subtypes may be associated with different degrees of protection. Diabetes 46:143-149

Vafiadis P, Bennett ST, Todd JA, Nadeau J, Grabs R, Goodyer CG, Wickramasinghe S, Colle E, Polychronakos C (1997) Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. Nat Genet 15:289-292

Vandiedonck C, Beaurain G, Giraud M, Hue-Beauvais C, Eymard B, Tranchant C, Gajdos P, Dausset J, Garchon HJ (2004) Pleiotropic effects of the 8.1 HLA haplotype in patients with autoimmune myasthenia gravis and thymus hyperplasia. Proc Natl Acad Sci U S A 101:15464-15469

Vandiedonck C, Capdevielle C, Giraud M, Krumeich S, Jais JP, Eymard B, Tranchant C, Gajdos P, Garchon HJ (2006) Association of the PTPN22*R620W polymorphism with autoimmune myasthenia gravis. Ann Neurol 59:404-407

Vang T, Congia M, Macis MD, Musumeci L, Orru V, Zavattari P, Nika K, Tautz L, Tasken K, Cucca F, Mustelin T, Bottini N (2005) Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. Nat Genet 37:1317-1319

Viken MK, Amundsen SS, Kvien TK, Boberg KM, Gilboe IM, Lilleby V, Sollid LM, Forre OT, Thorsby E, Smerdel A, Lie BA (2005) Association analysis of the 1858C>T polymorphism in the PTPN22 gene in juvenile idiopathic arthritis and other autoimmune diseases. Genes Immun 6:271-273

Vincent A, Palace J, Hilton-Jones D (2001) Myasthenia gravis. Lancet 357:2122-2128

Vaaler S, Møinichen T, Grendstad I (2004) Diabeteshåndboken. Gyldendal Akademiske

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434:338-345

Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30:3754-3766

Zavasnik-Bergant T, Turk B (2006) Cysteine cathepsins in the immune response. Tissue Antigens. In press.

Zheng W, She JX (2005) Genetic association between a lymphoid tyrosine phosphatase (PTPN22) and type 1 diabetes. Diabetes 54:906-908

Zhernakova A, Eerligh P, Wijmenga C, Barrera P, Roep BO, Koeleman BP (2005) Differential association of the PTPN22 coding variant with autoimmune diseases in a Dutch population. Genes Immun 6:459-461

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89-100

Zuklys S, Balciunaite G, Agarwal A, Fasler-Kan E, Palmer E, Hollander GA (2000) Normal thymic architecture and negative selection are associated with Aire expression, the gene defective in the autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED). J Immunol 165:1976-1983

Aaltonen J, Björses P, Perheentupa J, Horelli−Kuitunen N, Palotie A, Peltonen L, Lee Y, Francis F, Henning S, Thiel C, Leharach H, Yaspo M (1997) An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. The Finnish-German APECED Consortium. Autoimmune Polyendocrinopathy-Candidiasis-Ectodermal Dystrophy. Nat Genet 17:399-403

Association of the *CTSL2* gene with type 1 diabetes and myasthenia gravis

Hege Dahlen[1], Marte K. Viken[1], Eva Tolosa[2], Geir Joner[3], Knut Dahl-Jørgensen[3], Dag E. Undlien[4], Erik Thorsby[1], Benedicte A. Lie[1]

[1] Institute of Immunology, Rikshospitalet-Radiumhospitalet Medical Center, University of Oslo, Norway
[2] Department of Neurology, Tübingen University Hospital, Tübingen University, Tübingen, Germany
[3] Diabetes Research Center – Aker and Ullevål University Hospitals, Department of Pediatrics, Ullevål Hospital, Oslo, Norway
[4] Institute of Medical Genetics, Faculty Division Ullevål University Hospital, University of Oslo, Oslo, Norway

The cathepsin L2 (*CTSL2*) gene, encoding a cysteine cathepsin involved in degradation of Invariant chain, was investigated as a candidate gene for type 1 diabetes (T1D) and myasthenia gravis (MG). A role in diabetes development in non-obese diabetic mice has been suggested, and discrepancy in expression between MG patients and controls has been observed. Four single nucleotide polymorphisms (SNP) and two microsatellites were genotyped in 429 Norwegian T1D trios, in addition to 83 German MG patients and 244 German controls. A haplotype consisting of two SNPs (rs10739289 and rs7875800) and one microsatellite (D9S971) was found to be associated with T1D (34T vs. 13NT, p=0.002). Another SNP, rs4743056, together with D9S971, showed association with MG (14.3% vs. 7.5%, p=0.02), and an even stronger association with the early onset subgroup of MG (25% vs. 7.5%, p=0.007). *CTSL2* was extensively tested for alternative splice variants and several different mRNA transcripts were identified. Relative mRNA expression of *CTSL2* was measured in 42 thymic tissue samples and correlated with the disease associated haplotypes, but no significant correlation was seen. Given the limited polymorphisms studies, the primary association has probably yet to be identified. Altogether, we have sustained *CTSL2* as a good candidate gene for autoimmune diseases in general.

Introduction

Autoimmune diseases, which affect approximately 3-5% of the population (Jacobson et al. 1997), are characterized by immune-mediated tissue destruction due to loss of self-tolerance. Most autoimmune diseases are multifactorial, and in general both genes and environment contribute to disease susceptibility. Studies have shown that areas linked to one autoimmune disease often overlap with areas linked to other autoimmune diseases (Becker et al. 1998). Thus, genes predisposing to autoimmune diseases can roughly be divided into two classes; those predisposing to autoimmune diseases in general, i.e. involved in the autoimmunity process, and those only predisposing to a specific disease, maybe directing the organ specificity (Becker 1999).

Type 1 diabetes (T1D) and myasthenia gravis (MG) are examples of complex autoimmune diseases, and both are characterized by the production of autoantibodies. In T1D, cytotoxic T cells (CD8+) destruct the insulin-producing pancreatic β-cells as a consequence of activation by T helper cells (CD4+) (reviewed in Tisch and McDevitt 1996). Autoantibodies directed against pancreatic components, like insulin and glutamic acid decarboxylase, are observed. In MG, the production of autoantibodies is directed against the acetylcholine receptors (AChR) of the neuromuscular junction. 85% of the patients have these antibodies, which leads to a blockage of signalling trough the receptor (Engel 1984; Drachman 1994; Vincent et al. 2001).

The cathepsin L2 (*CTSL2*) gene is shown to be expressed in cortical thymic epithelial cells (cTEC) (Tolosa et al. 2003). cTECs are known to be involved in positive selection of T lymphocytes in the thymus, and positive selection has been recognized, in addition to negative selection, to be important in establishment of self-tolerance (Kretz-Rommel and Rubin 2000). In 2003, Tolosa and colleagues observed higher expression of *CTSL2* in thymi from MG patients with thymoma or thymitis when compared with controls (both healthy and patients with thymoma, but without MG), and they suggested *CTSL2* to be involved in the balance between positive and negative selection. Several cathepsins are involved in the degradation of invariant chain (Ii), and thereby antigen presentation on HLA class II molecules. *CTSL2* is homolog to the murine cathepsin L, which has a unique role in the cortical part of the murine thymus, where positive selection takes place (Nakagawa et al. 1998; Honey et al. 2002). Furthermore, the murine cathepsin L gene has been suggested to play a role in development of diabetes in non-obese diabetic (NOD) mice (Maehr et al. 2005). In a genome wide screen for linkage in Scandinavia the gene area near *CTSL2* was found to be linked to T1D (Nerup and Pociot

2001). Based on this knowledge, we propose *CTSL2* to be a candidate gene for susceptibility to autoimmune diseases in general.

Accumulating evidence is pointing in the direction of regulatory polymorphisms, sometimes located far from the coding regions, being involved in predisposition to complex diseases (Pastinen et al. 2005; Stranger et al. 2005; Duff 2006). Polymorphisms in gene regulatory regions have previously been reported to be associated with autoimmune diseases. For instance, a variable number of tandem repeats (VNTR) upstream of the insulin gene is known to be associated with T1D (Bennett et al. 1995; Undlien et al. 1995). The protective variants have been found to increase the insulin mRNA expression in thymus and thereby possibly promoting negative selection of insulin autoreactive T cells (Pugliese et al. 1997; Vafiadis et al. 1997). Another example is *CTLA4*, where polymorphisms downstream of the gene have been found to increase the risk of T1D and Grave's disease (Ueda et al. 2003; Furugaki et al. 2004). The risk polymorphisms affects the mRNA splicing, and higher levels of a soluble variant is produced. Prolonged activation of T cells is then believed to take place, since there is less CTLA4 in the membrane to send inhibitory signals (Ueda et al. 2003).

In the present study, *CTSL2* was genetically investigated in both T1D and MG, with the help of six polymorphisms covering the gene locus. In addition, functional analysis of the gene was performed. The gene was extensively tested for alternative splice variants on thymic cDNA. Gene expression was analysed in 42 thymic samples and the expression levels were correlated with disease predisposing variants.

Materials and methods

DNA from 429 Norwegian trio families with one child diagnosed with T1D was used for the genotyping. All probands were diagnosed before the age of 15 and fulfilled the WHO criteria for T1D. DNA from 83 German MG patients together with 244 healthy German controls was also used for genotyping. MG patients were sub grouped into early onset myasthenia gravis (EOMG, before the age of 40; n=31) and late onset myasthenia gravis (LOMG, after the age of 40; n=42). MG patients were diagnosed according to the MGFA Classification (Jaretzki et al. 2000), and approximately 90% were AChR antibody positive.

A commercially available thymic total RNA sample (BD Biosciences Clontech, San Jose, CA 95131, USA) pooled from three Caucasian adult individuals (age 20-38) was used for investigation of alternative splice variants of *CTSL2* in thymus (hereafter referred to as adult thymic sample). In addition, total RNA from one fetal thymic tissue sample was used (aborted between week 20-26; referred to as fetal thymic sample). RNA from thymic tissue samples from 42 Norwegian children (22 female/20 male) under the age of 13 (26 under 1 year; 16 between 1-13 years), undergoing corrective cardiac surgery, were used for gene expression analysis (referred to as juvenile thymic samples). DNA from these 42 thymic samples was also genotyped for the same polymorphisms as the T1D and MG materials.

Whole genome amplification of DNA was performed prior to genotyping (GenomiPhi^{TM} Amplification Kit, GE Healthcare, Little Chalfont, UK). This is a reliable and high quality method, and has no implication for the genotyping results (Barker et al. 2004).

The study was approved by The Regional Committee for Research Ethics and written informed consent had been obtained from all study participants.

Primer design. Primers were either designed by Applied Biosystems (Foster City, CA 94404, USA), or designed using the Primer 3 program (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) and potential self-annealing and hairpin formation within the suggested primers were checked by using the Oligonucleotide Properties Calculator program (www.basic.northwestern.edu/biotools/OligoCalc.html). All primers were checked for gene specificity using BLAST (http://ncbi.nlm.nih.gov/BLAST/). Primers were either purchased from Applied Biosystems or Eurogentec (4102 Seraing, Belgium). Fluorescent labels at 5'-ends for the microsatellite primers and gene expression probes were FAM^{TM}, and the allelic discrimination probes had either FAM^{TM} or VIC$^{®}$ labels. $MgCl_2$ concentration and annealing temperature were optimised for each primer pair. For primer information; see supplementary table 5.

Gel electrophoresis. Amplified PCR products were tested on 1-2% agarose gel (Sigma-Aldrich, St. Louis, MO 63103, USA) with 0.5 µg/mL Ethidium Bromide (Mercury, Pretech Instruments KB, 191 44 Sollentuna, Sweden) and 1xTBE-buffer. GeneRuler^{TM} 50 bp, 1 kb and ΦX DNA ladders were used to determine product length of the DNA fragments (Fermentas, Ontario L7N 3N4, Canada).

Marker selection. Based on comparison (http://pipeline.lbl.gov; http://genome.ucsc.edu/) of the gene area around *CTSL2* in different species (Homo sapiens, Mus muculus, Rattus rattus, Pan troglodytes and Canis familiaris), a selection of SNPs in conserved regions were made, since conserved sequences are more likely to be of functional importance (Loots et al. 2000). At the time of SNP selection, 7 validated *CTSL2* SNPs were deposited in the databases (dbSNP, build 121). SNPs with different allele frequencies were selected to maximize the chances of picking up association through linkage disequilibrium (LD) (Johnson et al. 2002). The following SNPs were selected: rs4743056 (T/C) situated 20 kb upstream of *CTSL2*, rs7875800 (A/G) in the sixth intron, rs15394 (A/C) in exon 8/3'UTR and rs10739289 (A/T) approximately 200 bp downstream of the gene. Two microsatellites flanking each side of *CTSL2* were also included: D9S1851 is located 223 kb downstream of *CTSL2*, while D9S971 is 365 kb upstream of the gene (Figure 1 and Supplementary table 5). Primers for genotyping microsatellites D9S1851 and D9S971 were selected from the GDB human genome database (www.gdb.org) and the primers were fluorescently labelled. The primers given in the database for D9S1851 were self-complementary and were redesigned using the primer design procedure described above.

Microsatellites. Microsatellites were genotyped using PCR and fragment length analysis. PCR was performed in 10 µL reactions containing 2mM $MgCl_2$ for D9S971 and 3 mM $MgCl_2$ for D9S1851, 2 µM of each primer and ~10 ng of template DNA. PCR conditions included an initial denaturation step at 94°C for 2 min, followed by 25 cycles of 94°C denaturation for 15 seconds, annealing (54°C for D9S971 and 61°C for D9S1851) for 30 seconds and elongation at 72°C for 1 min, and a final elongation step at 72°C for 5 min. Fragment lengths were analysed on an ABI3730 DNA Sequencer with GeneScan^{TM}-500 LIZ^{TM} Size Standard and POP-7^{TM} polymer (Applied Biosystems). GeneMapper software (Applied Biosystems v. 3.5 and 3.7) was used to analyse the genotyping results.

Both microsatellites consisted of a dinucleotide repeat. Twelve alleles were observed for D9S1851 in the T1D material, with the PCR products ranging from 220-246 bp. In the MG material an additional allele, 219 bp, was observed in one of the samples. For D9S971, ten different alleles ranging from 113-133 bp were observed in both materials. For presentation the alleles were renamed 1, 2, 3 etc starting with the shortest allele seen in the total data set.

Allelic discrimination. SNPs were genotyped using TaqMan® SNP Genotyping assays (Applied Biosystems). One of the assays was pre-designed by Applied Biosystem (By demand, prod. no: C___8737114_10), while the other assays were designed after submittance of sequences covering the SNPs (By design) (Supplementary table 5).

All reactions were run in 5 µL volumes with ~2.5 ng of template DNA, using 96 or 384 well plates and ABgene® 2x QPCR Master Mix (ABgene, Epsom KT19 9AP, UK). PCR conditions included an initial denaturation step at 95°C for 15 min, followed by 40 cycles of 95°C denaturation for 15 seconds and annealing/elongation at 60°C for 1 min. PCR was either performed on regular PCR block or as real-time on ABI7000 or ABI7900 Real-Time PCR System, but always followed by post-PCR reading on either ABI7000 or ABI7900. Allelic discrimination was analysed with SDS software (Applied Biosystem v. 1.1 and 2.2).

Isolation of RNA and cDNA synthesis. The juvenile thymic tissue samples were collected directly into RNAlater in appropriate sample sizes, with at least one side being less than 0.5 cm thick and 5 mL reagent per 1 gram, to protect RNA from degradation, and stored at -20°C as recommended by manufacturer (Ambion, Austin, TX 78744-1832, USA). Total RNA and DNA were extracted from approximately 100 mg thymic tissue with TRIzol® Reagent according to the manufacturer's protocol (Invitrogen, Carlsbad, CA 92008, USA). RNA samples were dissolved in nuclease free water, while for DNA, 1xTE-buffer was used. Final concentrations of total RNA were measured with BIORAD SmartSpec^{TM} 3000 at 260 nm, and DNA was measured with NanoDrop at 260 nm (NanoDrop Technologies, Wilmington, DE 19810, USA). RNA samples were stored at –70°C, while DNA samples were stored at – 20°C.

cDNA synthesis was performed from total RNA by reverse transcription using SuperScript^{TM} III Reverse Transcriptase (Invitrogen, Cat. no: 18080-051), random hexamers, and ~1000 ng of total RNA as input in each reaction. cDNA from the fetal thymic sample was supplied by Christopher Bowlus, and was prepared with the same procedure and reagents as described, except that the tissue was frozen in liquid Nitrogen prior to isolation of RNA.

Alternative splice variants/GeneRacer^{TM}. Two different strategies were applied in the search for alternative splice variants of *CTSL2* in thymic tissue. First, primers were designed based on mRNA information from Ensembl (http://www.ensembl.org/index.html, v.36). The primers were located in exon 1/5'UTR, exon 2 and exon 8/3'UTR (Figure 3A and Supplementary table 5). This method was applied to study all variants with divergent sequences between exon 1 or exon 2 and 3'UTR. The second strategy was to identify splice variants with the help of rapid amplification of cDNA ends (RACE) (GeneRacer^{TM}, Invitrogen). GeneRacer^{TM} is a RNA ligase mediated rapid amplification of 5' and 3' cDNA ends. Primers for GeneRacer^{TM} were designed as described by the manufacturer. Since both 5'RACE and 3'RACE were to be performed, gene specific primers were designed in both directions. In addition, nested primers were designed (Supplementary figure 1A and Supplementary table 5). Again, aberrant transcripts were to be studied, in addition to possible transcripts with variable length of the 5'- and 3'-ends.

In the first strategy, cDNA from the adult and fetal thymic samples were used. PCR was performed in both 20 and 40 µL reactions with 2-3 mM $MgCl_2$, 10 µM of each primer (Eurogentec) and 25-50 ng of template cDNA. PCR conditions included an initial denaturation step at 94°C for 10 min, followed by 40 cycles of 94°C denaturation for 45 seconds, annealing at 58°C for 30 seconds and elongation at 72°C for 2 min, and a final elongation step at 72°C for 10 min. PCR products were tested on agarose gel as described above.

The second strategy utilized total RNA from adult thymic tissue, in addition to one of the juvenile thymic samples. The integrity of RNA was checked by agarose gel electrophoresis, and 28S and 18S rRNA bands were observed. Total RNA was treated according to the GeneRacer^{TM} protocol and GeneRacer^{TM} RNA oligos were ligated to the mRNA ends (Invitrogen). cDNA synthesis was performed as described above. Subsequently, both 5' and 3'RACE were performed as described by the manufacturer. Annealing temperature used was 66°C. Elongation temperature was decided based on the Taq polymerase used; 5'RACE was performed with Platinum® *Taq* DNA polymerase High Fidelity (Invitrogen, elongation temperature 68°C), while 3'RACE and nested PCR's were performed with AmpliTaq Gold® DNA polymerase (Applied Biosystems, elongation temperature 72°C). Nested PCRs were performed for both 5'RACE and 3'RACE products as suggested by the manufacturer, with the following modifications; PCR reaction volume was decreased to 25 µL, $MgCl_2$ concentration changed to 1.5 mM, annealing temperature used was 65°C and 25 cycles were run. PCR products were tested on agarose gel as described above.

Cloning. PCR products were cut from agarose gel and purified by employing the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI 53711 USA). Cleaned PCR products were either cloned into pGEM®-T Easy TA Vector (Promega) or pCR® 2.1-TOPO® TA vector (Invitrogen). Both procedures were performed according to manufacturer's protocol. For pGEM®-T Easy Vector ligation was performed at 4°C over night, while for pCR® 2.1-TOPO® TA vector ligation was performed in room temperature for 20 min. The pCR® 2.1-TOPO® TA vector has a covalently bound Topoisomerase I, ensuring rapid ligation of the insert.

One Shot® TOP10 Chemically Compentent *E.coli* cells were transformed as described in the protocol (Invitrogen). Cells were plated onto LB-plates containing 50 or 100 µg/mL ampicillin, enabling selection of cells containing a plasmid. Positive and negative controls were applied as described in both protocols. For the pGEM®-T Easy Vector, LB-plates were prepared with IPTG (isopropyl thiogalactoside) and X-Gal (5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside). IPTG induce the activity of β-galactosidase which then hydrolyze X-Gal, resulting in a colour change from white to blue if no PCR product is inserted into the vector. Hence; white colonies have insert, blue colonies have no insert. In the pCR® 2.1-TOPO® TA vector, inserts interrupt the lethal *E.coli* gene *ccdB*, which is fused with the C-terminus of *lacZ* in the vector; hence cells survive. Cells not containing insert in the vector, are killed upon plating. In addition to selection as describes above, colonies were screened by PCR to make sure they had the right insert. Plasmids were then purified from over night cultures with Wizard® *Plus* Miniprep DNA Purification System (Promega).

Sequencing. Plasmid inserts were amplified with M13 forward and reverse primers (Invitrogen and Eurogentec) and the BigDye® Terminator Sequencing Kit v 3.1, with reduced amount of BigDye® Terminator reagent (reduced to 1/8) (Applied Biosystems). Reactions were run in 10 µL volumes. Products were precipitated with Ethanol and EDTA (Sigma), and further resolved in Hi-Di™ Formamide (Applied Biosystems). Sequences were run on an ABI3730 DNA Sequencer with POP-7™ polymer and 36 cm long capillaries (Applied Biosystems). Results were analysed with SeqScape v.2.5 (Applied Biosystems), BLAT (http://genome.ucsc.edu), BLAST (http://ncbi.nlm.nih.gov/BLAST/), ClustalX (http://bips.u-strasbg.fr/fr/Documentation/ClustalX/,v.1.81), Genedoc (http://www.psc.edu/biomed/genedoc/ v.2.6.002 , Nicholas et al. 1997) and the mRNA sequences were translated into protein using Expasy (http://au.expasy.org/tools/dna.html). All sequences were aligned to the human genome reference sequence and adult tissue mRNA showed 99.9% identity with *CTSL2* mRNA in the database, while fetal tissue mRNA showed 99.8% identity.

Gene expression analysis. An expression profile of *CTSL2* can be found at http://harvester.embl.de/harvester/O609/O60911.htm. Expression of *CTSL2* in thymus was determined using TaqMan® Gene Expression Assays (Part. no: 4331348, Applied Biosystems) on cDNA. *CTSL2* expression was measured in 42 juvenile thymic samples. DNA from the same samples was genotyped, as described above, prior to the investigations. Primers and probes for TaqMan® Gene Expression Assays were designed by Applied Biosystems after submission of sequences (Supplementary table 5). Three different assays covering exon-exon boundaries of *CTSL2* were designed (Supplementary figure 1B). Assay 1 (exon 1/2) represented all transcripts with exon 1/exon 2 boundary present. Assay 2 (exon 5/6) represented all transcripts with exon 5/exon 6 boundary present. Assay 3 (exon 6/7) represented all transcripts with exon 6/exon 7 boundary present. The housekeeping gene beta-2-microglobulin (B2M) was used as an endogenous control (TaqMan® Pre-Developed Assay Reagents, Human B2M, Part. no: 4333766F, Applied Biosystems). Reactions were run in 10 µL volume, and 4.5 ng cDNA pr reactions was used for assay 1-3, while 0.5 ng was used for the B2M assay. All samples were run in triplicates and outliers were removed during analysis. A cDNA standard curve was applied to obtain relative quantifications results. Real-time PCR was performed on ABI7900 Real-Time PCR System. PCR conditions included an initial step at 50°C followed by a denaturation step at 95°C for 10 min. Then 40 cycles of 95°C denaturation for 15 seconds and annealing/elongation at 60°C for 1 min were run. Results were analyzed with SDS2.2 software (Applied Biosystems).

Statistics. Power calculator located at http://pngu.mgh.harvard.edu/~purcell/gpc/ was employed for the T1D family material, while for the MG case-control material the power was calculated at http://calculators.stat.ucla.edu/powercalc (Table 1). Minor allele frequencies (MAF) were found in the dbSNP database (build 121, http://www.ncbi.nlm.nih.gov/projects/SNP/) prior to the study. Due to inconsistent MAF values in the database, the observed MAF values were used post hoc.

Mendelian inheritance of the polymorphisms was checked in the T1D family material, in order to detect marker typing incompatibilities (genotyping errors) by using the program PedCheck (O'Connell and Weeks 1998). Hardy-Weinberg equilibrium (HWE) was tested for all polymorphisms in both materials, and deviation from HWE would suggest genotyping errors (Hosking et al. 2004). HWE for the microsatellites were calculated with the Arlequin software in the MG material (http://anthro.unige.ch/arlequin/, v.2.000; (Schneider et al. 2000) and Pedstats in the T1D family material (Wigginton and Abecasis 2005), while Haploview was used for the SNPs in both material (http://www.broad.mit.edu/mpg/haploview/, v.3.2; (Barrett et al. 2005).

Single point and multipoint association analyses were performed using the Unphased package (http://portal.litbio.org/). The Tdtphase application was

Table 1 Estimation of the risk (given by GR for T1D and OR for MG) that each polymorphism must exercise in order for our data sets to have >80% power to detect an effect with statistical significance p<0.05.

	dbSNP, a priori				dbSNP, post hoc				
Polymorphism	MAF[1]	T1D (GR)	MG (OR)	EOMG (OR)	MAF[2]	T1D (GR)	MAF[3]	MG (OR)	EOMG (OR)
rs10739289	0.150	1.6	2.4	3.5	0.12	1.6	0.14	2.5	3.6
rs15394	0.468	1.9	2	3.1	0.13	1.6	0.14	2.5	3.6
rs7875800	0.320	1.6	2	3.0	0.13	1.6	0.14	2.5	3.6
rs4743056	0.495	1.9	2	3.1	0.47	1.9	0.45	2.1	3.0

MAF – minor allele frequency; [1] – MAF value from dbSNP build 121; [2] – observed MAF values in the T1D family material; [3] – observed MAF values in the MG material; GR – genotype risk; OR – odds ratio

used for the family-based tests, while the Cocaphase application was used for the case-control material. Association was tested in the T1D families by both the transmission disequilibrium test (TDT) and the affected family-based control test (AFBAC). TDT considers transmission of alleles from heterozygous parents to affected offspring (proband) (Spielman et al. 1993), while the AFBAC test includes transmissions from all parents, not only from the heterozygous (Thomson 1995). In the MG case-control material allele or haplotype frequencies were compared using chi-square statistics, or Fisher exact test when appropriate.

Haplotype frequencies were estimated using the Expectation-Maximization (EM) algorithm (Excoffier and Slatkin 1995) implemented in the cocaphase program for the MG data set, and for the AFBAC analysis for T1D. Only phase-certain haplotypes were included in the TDT analyses for the T1D data set. p-values <0.05 were considered statistical significant. p-values are presented as non-corrected, but all statistical significant p-values are also given as corrected values. For the microsatellites only alleles with a frequency >1% were tested. For T1D, 95% confidence intervals (CI) for the percentage transmission (%T) of alleles were calculated by application of the formula: CI= %T ±1.96 $\sqrt{(p(1-p)/n)}$, where p=proportion of positive transmissions and n denotes the total number of transmissions. For the MG material, odds ratio (OR) and 95% CI values were calculated by Statcalc Epi.v.5 (Dean et al. 1991).

LD between the polymorphisms was calculated as D' and r^2 using the Unphased package and/or Haploview v.3.2. Originally Haploview can only be applied on SNPs (biallelic polymorphisms), therefore the multiallelic microsatellites had to be changed into biallelic markers, by addressing one of the allele as 1, and all other alleles as 2.

Gene expression data was normalized with respect to the endogenous control B2M using the following calculations:

$$\text{Relative mRNA expression per sample} = \frac{[\text{Quantity mean value from assay (Ex: Exon 1/2)}]}{[\text{Quantity mean value from B2M assay}]}$$

Gene expression results were then grouped according to the associated haplotypes in T1D and MG, and Graphpad Prism 4 was used to visualize the gene expression data (http://www.graphpad.com/, v.4.03).

Results

In order to test *CTSL2* as a candidate gene for T1D and MG, four SNPs and two microsatellites (Figure 1) were genotyped in 429 Norwegian T1D trio families, 83 German MG patients and 244 German controls. The genetic polymorphisms were in HWE in all datasets studied, and few Mendelian errors (< 3%) were observed in the T1D family data set. Prior to analysis the Mendelian errors were dissolved by removing the genotypes for the relevant polymorphisms for the families in question.

Association with *CTSL2* polymorphisms detected in both T1D and MG. Results from the single locus association analyses are shown in Table 2 and Table 3. Microsatellite D9S971 showed significant association with T1D and the EOMG subgroup, but not with the total MG data set. Sub grouping of the T1D material by age was not needed, since all T1D cases were diagnosed before the age of 15. In T1D, allele 8 at D9S971 was more often transmitted to the affected offspring (proband) than the expected 50% (67% T, p=0.01), and this association was also seen with AFBAC analysis (4.5% vs. 2.2%; p=0.01). In addition, two of the SNPs (rs10739289 and rs7875800) showed a biased transmission (57% T), however this did not reach statistical significance (p=0.07).

As previously mentioned, microsatellite D9S971 was also associated with EOMG, however the association was with allele 4 (OR=1.95, p=0.02) and not allele 8 as in T1D. Furthermore, allele C at rs4743056 was found more often in the EOMG patients than controls, but did not reach statistical significance (OR=1.6, p=0.08). Notably, the frequency distribution for the two SNPs (rs10739289 and rs7875800) showing

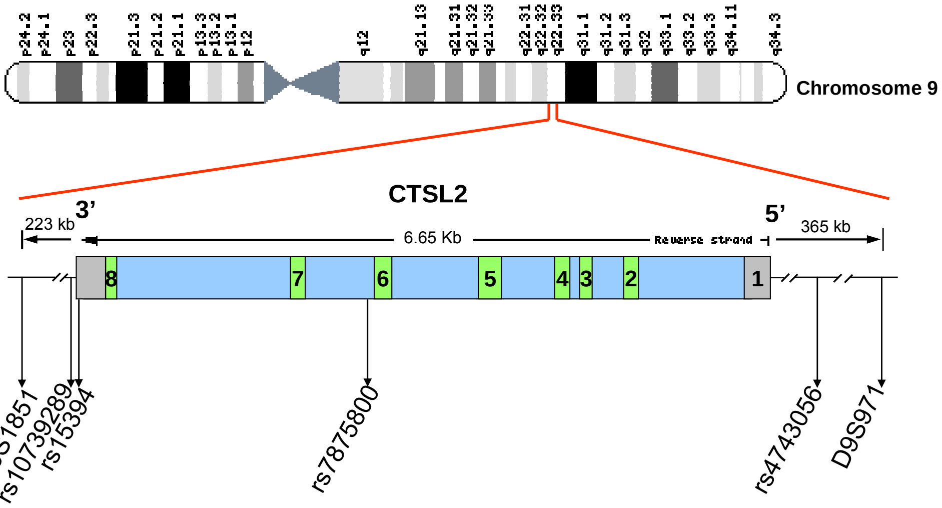


Figure 1 Schematic drawing of the *CTSL2* gene indicating the positions of polymorphisms. *CTSL2* is located on chromosome 9 (9q22.2). Microsatellite D9S1851 is situated 223 kb downstream; rs10739289 (A/T) approximately 223 bp downstream; rs15394 (A/C) in exon 8/3'untranslated region (UTR); rs7875800 (A/G) in the sixth intron; rs4743056 (T/C) 20 kb upstream; D9S971 is 365 kb upstream.

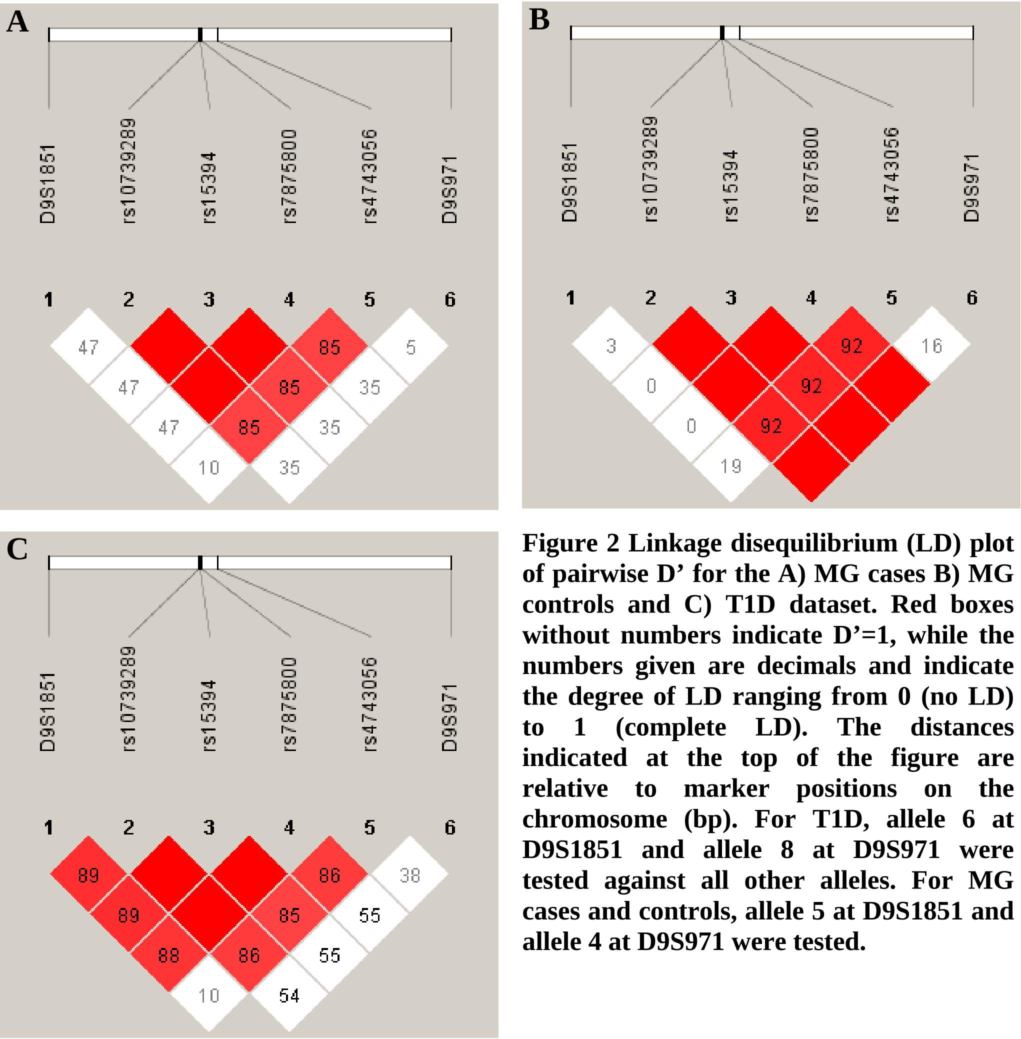


Figure 2 Linkage disequilibrium (LD) plot of pairwise D' for the A) MG cases B) MG controls and C) T1D dataset. Red boxes without numbers indicate D'=1, while the numbers given are decimals and indicate the degree of LD ranging from 0 (no LD) to 1 (complete LD). The distances indicated at the top of the figure are relative to marker positions on the chromosome (bp). For T1D, allele 6 at D9S1851 and allele 8 at D9S971 were tested against all other alleles. For MG cases and controls, allele 5 at D9S1851 and allele 4 at D9S971 were tested.

Table 2 Association analyses of the *CTSL2* polymorphisms in the T1D trio material by the transmission disequilibrium test (TDT) and affected family-based control test (AFBAC)

		TDT					AFBAC			
Polymorphism	Allele	T	NT	%T	95% CI	p-value	AFBAC-T% n = 858	AFBAC-NT% n = 858	OR (95% CI)	p-value
D9S1851	6	203	175	53.7	49-59	0.15	39.2	35.5	1.17 (0.94-1.44)	0.14
rs10739289	T	90	67	57.3	50-65	0.07	89.9	87.0	1.33 (0.97-1.85)	0.07
rs15394	A	92	71	56.4	49-64	0.10	89.5	86.9	1.29 (0.94-1.78)	0.10
rs7875800	A	97	73	57.1	50-64	0.07	89.4	86.4	1.33 (0.97-1.82)	0.06
rs4743056	T	196	180	52.1	47-57	0.41	54.0	51.9	1.09 (0.89-1.33)	0.42
D9S971	8	35	17	67.3	55-80	0.01*	4.50	2.20	2.11 (1.13-3.96)	0.01*

The most positively associated allele was used for each polymorphism; T – number of transmitted alleles from heterozygous parents; NT – number of non-transmitted alleles from heterozygous parents; %T – percent transmission (deviation from 50% indicate association); 95% CI – 95% confidence interval for the %T, AFBAC-T – allele frequency among cases; AFBAC-NT – allele frequency among family-based controls; n – total number of alleles; OR (95% CI) – odds ratio with 95% confidence interval; Analysis were performed with 429 T1D families; p-values are uncorrected; * corrected p-value=0.18

Table 3 Association analyses of the *CTSL2* polymorphisms in the MG case-control material.

Polymorphism	Allele	All MG cases (%) N = 83	EOMG cases (%) N = 31	LOMG cases (%) N = 42	CTR (%) N = 244	OR (95% CI) all MG	p-value all MG	OR (95% CI) EOMG	p-value EOMG	OR (95% CI) LOMG	p-value LOMG
D9S1851	5	31.3	30.65	29.8	25.0	1.37 (0.91-2.05)	0.12	1.33 (0.71-2.45)	0.35	1.27 (0.74-2.18)	0.36
rs10739289	T	86.1	88.71	85.7	85.9	1.02 (0.59-1.75)	0.95	1.28 (0.54-3.23)	0.54	0.98 (0.49-2.01)	0.95
rs15394	A	86.1	88.71	85.7	86.3	0.99 (0.58-1.71)	0.97	1.25 (0.52-3.15)	0.59	0.96 (0.47-1.97)	0.90
rs7875800	A	86.1	88.71	85.7	86.2	1.00 (0.58-1.72)	0.99	1.26 (0.53-3.17)	0.57	0.96 (0.48-1.98)	0.91
rs4743056	C	52.4	56.45	51.2	44.6	1.37 (0.95-1.98)	0.08	1.61 (0.91-2.83)	0.08	1.30 (0.85-2.12)	0.27
D9S971	4	28.8	36.67	22.9	22.9	1.36 (0.89-2.07)	0.14	1.95 (1.06-3.55)	0.02*	0.98 (0.53-1.78)	0.93

The most positively associated allele was used for each polymorphism; MG – myasthenia gravis; EOMG – early onset MG; LOMG – late onset MG; CTR – controls; N – total number of cases/controls; OR (95% CI) – odds ratio with 95% confidence interval; p-values are uncorrected; * corrected p-value=0.8

association with T1D (Table 2), had equal distribution in the EOMG material (Table 3), approximately 89% in cases and 86% in controls.

LD analysis was performed for both materials separately, and complete LD was seen between the SNPs rs10739289, rs15394 and rs7875800 in both materials (D'=1, Figure 2 A, B and C). Furthermore, rs4743056 showed strong LD with the three mentioned SNPs in both T1D and MG as well, although this was not complete (Figure 2 A, B and C). The microsatellites were converted to biallelic markers for these analyses, with the most biased alleles tested against all others (i.e. D9S1851-allele 6 and D9S971-allele 8 in T1D, and D9S1851-allel 5 and D9S971-allele 4 in MG). There was some degree of LD between the microsatellites and SNPs rs10739289, rs15394 and rs7875800 for both materials, but keep in mind that the alleles tested were different in the two materials. In T1D, the associated allele 8 at D9S971 showed an LD ranging from D'=0.54-0.55 with these SNPs (Figure 2 C). In the MG cases, the disease associated allele 4 at D9S971 showed an LD of D'=0.35 with the mentioned SNPs, while for the controls complete LD was observed (Figure 2 A and B).

Next we investigated the associations with haplotypes in both materials. For T1D, the different haplotype constellation can be viewed in (Supplementary table 2). We investigated the borderline associated rs10739289 (T) and rs7875800 (A) as a haplotype, and the T-A haplotype achieved significant association with T1D (59%T, p=0.04). Further investigations showed that this two-SNP haplotype, together with the predisposing allele 8 at microsatellite D9S971, was even stronger associated (72%T, p=0.002). In MG, a possible haplotype between rs4743056 (C) and D9S971 (4) was investigated (data not shown). The analysis revealed the C-4 haplotype to be associated with MG (OR=1.86, p=0.02). Once more the association seemed to follow the EOMG subgroup (OR=2.24, p= 0.0008) and not the LOMG subgroup (OR=1.51, p= 0.54). After correction of the p-value, the association seen with EOMG was still significant (p=0.032).

A second LD test was applied to further reveal the relationship between the T1D associated allele 8 at D9S971 and the three SNPs that showed a biased transmission in T1D. rs7875800 represented all three SNPs in this test, as the SNPs were in complete LD (Figure 2 C). The alleles transmitted to the T1D probands showed much stronger LD than the non-transmitted alleles (which represents the family-based controls) (D'=1 and r^2=0.009 vs. D'= -0.037 and r^2=0.0002). Similar LD analysis was performed for the associated haplotype in the MG material, however no LD was observed between allele C at rs4743056 and allele 4 at D9S971 in neither patients (D'= -0.053, r^2=0.001) nor controls (D'= -0.025, r^2=0.015).

Genotypic association analyses were performed to further study possible dose effects of the associated variants on disease risk. For T1D, the genotypic analyses for the two-SNP haplotype (rs10739289-rs7875800) revealed statistical significant association (p = 0.002, corrected p-value=0.04, Supplementary table 3). Further analyses showed that the association represented a recessive risk model, where one needs to be homozygous for the associated haplotype to have increase susceptibility (OR=1.9, p=0.0006, corrected p-value=0.01, Supplementary table 3). In the MG material, the associated haplotype was statistically significant for the EOMG subgroup (p=0.003, corrected p-value=0.1, Supplementary table 4). The frequency distribution between cases and controls suggested a dominant risk model (OR=3.6, p=0.0007, corrected p-value=0.03, Supplementary table 4).

Several alternative *CTSL2* transcripts are expressed in thymus. To study the possible functional implications of the genetically associated polymorphisms, we wanted to investigate expression of *CTSL2* in thymic tissue. Initially, we searched for alternative variants using two different strategies. First, transcripts of *CTSL2* between two pre-defined borders were studied, and although no alternative variants were found, expression of *CTSL2* mRNA was observed in both adult and fetal thymic tissue (Figure 3A). Two different forward primers in exon 1 were used (termed A and B) against one 3'UTR reverse primer, but the exon 1A primer did not give amplification product (not shown). Comparisons of the *CTSL2* mRNA sequence from different databases revealed that this primer might have been located in an untranscribed part of 5'UTR/exon 1. Both Ensembl and NCBI reported longer mRNA sequence than VEGA (1547 bp and 1496 bp vs. 1369 bp, respectively). The main difference was in the length of 5'UTR/exon 1, which was reported to be 240 bp by Ensembl, 187 bp by NCBI and 63 bp by VEGA. The mRNA information from VEGA therefore seemed consisted with our results. The second forward primer, exon 1B, was moved 44 bp further downstream in exon 1/5'UTR, and gave amplification of *CTSL2* mRNA in both adult and fetal thymic tissue.

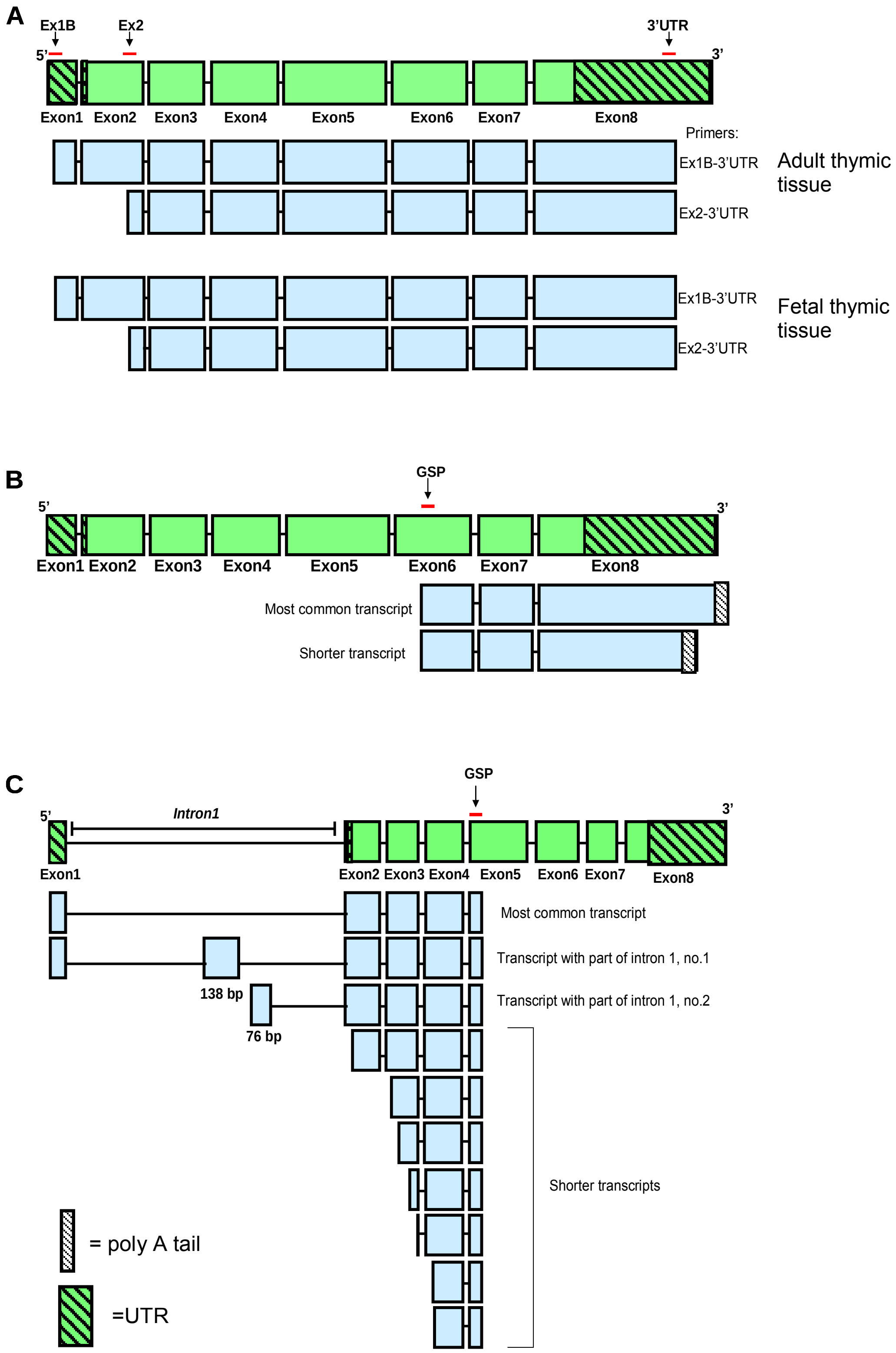


Figure 3 cDNA sequences from fetal, juvenile and adult thymic tissue. Green boxes indicate the exons of full length (1369 bp) reference sequence of *CTSL2* mRNA from VEGA (http://vega.sanger.ac.uk); Blue boxes represent our sequencing results from; A) adult and fetal thymic tissue using PCR amplification, B) RACE analysis of the 3' end (juvenile and adult thymic tissue); C) RACE analysis of the 5' end (juvenile and adult thymic tissue); Primer positions are marked red; GSP – gene specific primer; UTR – untranslated region

Secondly, the 5'- and 3'-ends of *CTSL2* present in thymus were studied, and several transcripts deviating from full length mRNA (1369 bp, VEGA) were identified. The limiting boundaries were set inside the gene as defined by gene specific primers (Supplementary figure 1A), and transcripts were sequenced to the very ends of both 5' and 3' UTRs (Figure 3B and C). 3'RACE identified a variant with shorter 3'UTR sequence, while 5'RACE identified several different transcripts. PolyA tail was identified in all 3' end sequences, but polyA-signal (aataaa) was only identified in the longer variants. Most transcripts detected were identical to the 5'- or 3' ends of the full length mRNA sequence (1369 bp, VEGA). In the majority of the cloned 5' end transcripts the reported start codon, 10 bp into exon 2, was identified.

Among the transcripts identified with 5'RACE, two had different parts of intron 1 included (Figure 3C). In the first variant, 138 bp of intron 1 was kept between exon 1 and 2 and the known start codon in exon 2 was identified. Another start codon was identified in the very end of the intron sequence, but this was in a different reading frame and led to a stop codon after just 25 amino acids. The second variant started inside intron 1, but no start codon was detected in front of the known start codon in exon 2. Several shorter variants were also detected with 5'RACE, and most had start codons in the same reading frame as the full length mRNA reported in VEGA.

No correlation between gene expression and disease associated haplotypes. In addition, *CTSL2* expression in thymic tissue was studied and correlated with the genetically associated polymorphisms. No significant differences were observed between relative gene expression and the T1D associated polymorphisms (Figure 4A). The results were grouped according to the statistically significant recessive model (Supplementary table 3); i.e. homozygous presence of the two-SNP haplotype vs. heterozygous or no presence. For the MG associated variants, gene expression results were correlated with the statistically significant genotypes following a dominant model (Supplementary table 4), but no significant differences were observed (Figure 4B).

Discussion

Here, we report an association between *CTSL2* and two autoimmune diseases; T1D and MG. This is the first report of a genetic association seen with the *CTSL2* gene, and might lead to the discovery of yet another disease predisposing gene. T1D and MG are both complex autoimmune diseases, and the sharing of risk genes is not surprising. Earlier both *CTLA4* (Huang et al. 1998; Ueda et al. 2003) and *PTPN22* (Bottini et al. 2004; Vandiedonck et al. 2006) have been shown to be associated with these diseases. The gene area near *CTSL2* has previously been linked to T1D through a genome-wide screen for linkage in Scandinavia (Nerup and Pociot 2001), and this support our findings in the T1D material. No such genome-wide scan has been performed in MG, but earlier reports have shown that gene areas linked to one autoimmune disease often are found to overlap with areas linked to other diseases (Becker et al. 1998). Expression of *CTSL2* was localized to cTECs (Tolosa et al. 2003), and these cells are known to be involved in positive selection in the thymus. Tolosa and colleagues (2003) showed higher expression in MG patients when compared with controls, and this further substantiate our finding in MG.

A critical step when studying a candidate gene is to ensure that all genetic variation in the gene is picked up either directly or indirectly. At the time of SNP selection for the present study, few SNP were validated by genotype frequencies in the *CTSL2* gene region (dbSNP, build 121). Validated SNPs were selected based on the allele frequencies reported at the time, since different allele frequencies often represent different LD patterns (Johnson et al. 2002). After the study was completed, updated allele frequencies were given in dbSNP (build 125). As it turned out, three of the tested SNPs (rs10739289, rs15394 and rs7875800) were in complete LD, and the allele frequencies showed discrepancies from the originally reported values. Because of this, the gene area was not genetically covered as well as intended. The observed association might not be the primary, and further studies including more SNPs should be performed in order to map the area more comprehensively.

We also aimed at selecting SNPs that might be functionally important. SNPs were selected based on their positions relative to the gene, and potential *cis*-acting polymorphisms, e.g. SNPs located in conserved regions, were selected (Figure 1); 3'UTR, intron, upstream and downstream of the gene. No SNPs in coding regions (exons) were known at the time of SNPs selection, however later a non-synonymous SNP in exon 4 was reported, although it is not validated to be a true polymorphism.

The microsatellite, D9S971, was single-handedly significantly associated with both T1D and the EOMG subgroup of the MG material. Different alleles at D9S971 were involved, which

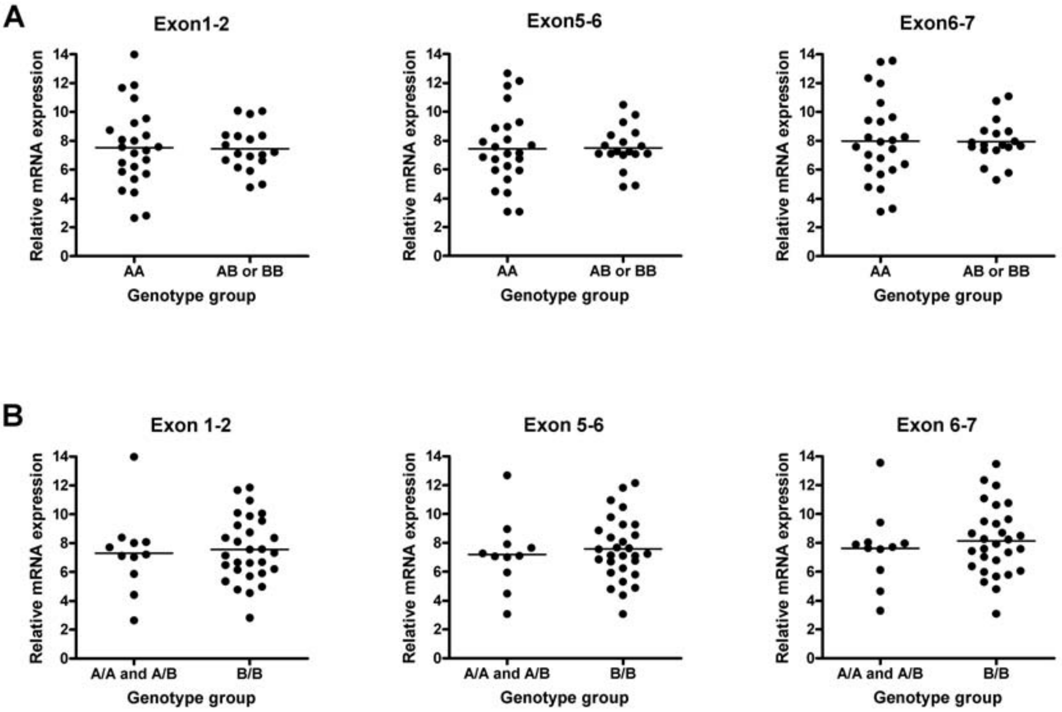


Figure 4 Relative mRNA expression of *CTSL2* in thymic tissue correlated with the genotypes associated with T1D (panel A) and MG (panel B). Three different gene expression assays were analysed for both materials, assay 1 (exon 1-2), assay 2 (exon 5-6) and assay 3 (exon 6-7); Panel A) T1D associated two SNP-haplotype based on the recessive disease model; A – presence of the two-SNP associated haplotype (rs10739289-allele T and rs7875800-allele A); B – lack of the T1D associated haplotype. Panel B) MG associated haplotype based on the dominant disease model; A – presence of MG associated haplotype (rs4743056-allele C and D9S971-allele4); B – no presence of MG associated haplotype.

might be due to the fact that we are studying two different autoimmune diseases, with different pathogenesis. Furthermore, the disease predisposing allele is not necessarily the same in T1D and MG. Studies in monogenic diseases, like cystic fibrosis (CF), have revealed that different polymorphisms in the same gene can give rise to the same disease (Kerem et al. 1989), and this might apply to complex phenotypes, like autoimmunity, as well. In addition, the disease materials were from two different demographic populations, and throughout time these populations might have undergone independent recombination in the genetic area studied. In contrast to the SNPs, which are close to or within the gene, D9S971 is 365 kb upstream of *CTSL2*. The further away a polymorphism is to the actual disease predisposing allele, the more likely it is that the gene area surrounding the studied polymorphism has been subject to recombination over time. The differences seen in LD values between the SNPs and the microsatellite observed in the two disease materials could also indicate the occurrence of differential recombination. An interesting observation was that rs10739289, rs15394 and rs7875800, showed the same imbalance in allele frequencies in both T1D and EOMG; 89% in cases vs. 86% in controls (Table 2 (AFBAC) and Table 3). The reason why the T1D associated SNPs are not significantly associated in EOMG might be due to the small sample size (N=31), and as the power calculation showed; the EOMG material is not large enough to pick up effects in this range. The similar imbalance seen might though indicate that the SNPs are closer to the causal disease allele, than the microsatellite D9S971. On the other hand, rs4743056, which is part of the MG associated haplotype, did not contribute to disease association in T1D. rs4743056 seems to be more heterogeneous than the other SNPs in the two populations, as the frequency differs largely in the control populations (44.6% in German controls vs. 51.9% in Norwegian family-based controls). This heterogeneity could again influence the LD pattern and thereby the ability to pick up an association. Altogether, the fact that two different diseases show association with the gene area studied strongly supports the probability of finding a true disease predisposing polymorphism in this area.

The reason why the EOMG, and not the total MG material or the LOMG subgroup showed association with *CTSL2,* might be explained by the great heterogeneity seen among MG patients. The EOMG and LOMG groups represent patients with different phenotypes; EOMG typically represents patients with thymic hyperplasia, presence of anti-AChR antibodies and a strong female bias, while LOMG represents patients with non-thymoma, presence of anti-AChR antibodies (but lower concentrations than in the EOMG subgroup), often thymus atrophy, and an equal occurrence in males and females (reviewed in Romi et al. 2005). The EOMG group might be more comparable with the T1D patients, since all diabetic patients in this study were diagnosed before the age of 15. The genetic contribution to disease development might be higher at lower age of onset, while the environmental contribution might dominate at later ages. Supporting this is earlier studies where the genetic association with HLA is shown to be stronger in early onset T1D (Caillat-Zucman et al. 1992). Interestingly, the MG subgroups show different HLA associations (reviewed in Romi et al. 2005), and this further supports the likelihood of observing deviating genetic associations in MG populations. Another interesting observation is that the EOMG group has a high frequency of concomitant autoimmune diseases (Thorlacius et al. 1989; Oosterhuis 1997), hence the probability of finding association with shared autoimmune predisposing genes is possibly larger in this group.

Further analysis revealed a recessive disease model to be statistically significant for the T1D material, while in MG a dominant model was suggested. However, in MG few individuals were homozygous for the associated polymorphisms, thus the groups were quite small, and thereby subject to uncertain estimates. Both diseases associated models need to be reconsidered when further studies to identify the primary involved polymorphisms have been conducted.

The fact that an association was seen in both T1D and MG supports *CTSL2* as a candidate gene for autoimmune diseases in general. The p-values for our reported associations were significant after correction for multiple testing, but nonetheless; to exclude that our finding is a false positive association, replication in other populations is needed, and if confirmed, the gene should be tested in other autoimmune diseases as well. The effect of *CTSL2* variants on disease risk cannot be reliably assessed before the directly involved variants are identified. The SNPs genotyped are not necessarily in complete LD with the primary association, and this would probably deflate our OR values. Nevertheless based on other known predisposing genes, like *PTPN22,* which have an OR = ~1.7 in T1D (Concannon et al. 2005), we expected *CTSL2* to exercise an effect in the observed range. It is important to recognize the impact even small effects can have on the phenotype, and often it is the involvement from several genes that together contributes to diseases development. Hence it is not necessarily the most statistically significant effect which is the most biologically important one (Stranger et al. 2005).

When genetically mapping a candidate gene, interesting results should be supplanted by functional studies. Understanding the effects of diseases associated polymorphisms on gene expression can provide important insight into why some polymorphisms are risk factors (Pastinen et al. 2006). Before embarking on expression analysis, identification of alternative transcripts of *CTSL2* in thymus was performed. Several alternative transcripts were identified in adult and juvenile thymic samples by employing two complementary strategies. The first strategy was limited by two defined outer boarders (in exon 1 and 3'UTR), and no alternative transcripts were identified, which indicated that no alternative variants with missing exons, or otherwise deviating sequences between exon 1 and 3'UTR, were expressed. On the other hand, we cannot exclude that alternative transcripts might exist in low copy number, making them hard to detect, or they might be individually expressed. *CTSL2* was nonetheless successfully observed in both adult and fetal thymic samples by sequencing, and expression of full length transcripts from exon 1 to 3'UTR was identified in both samples.

The second strategy (GeneRacer^{TM}) identified several transcripts with variable length in the ends, one variant in the 3' end and several in the 5' end (Figure 3B and C). Due to the nature of the GeneRacer^{TM} method, the identified transcripts should not have been subject to degradation. None of the observed variants could possibly have been identified with the first strategy, due to the primer positions, except for the transcripts identified with parts of intron 1 included between exon 1 and exon 2. The reason for not identifying these transcripts with the first method might, as previously mentioned, be low copy number or only expression in few individuals. It is important to point out that the 5'- and 3' ends are studied separately when applying the GeneRacer^{TM} method, hence no conclusions regarding the relationship between the different identified 5' and 3' transcripts can be drawn. The variants with parts of intron 1 included, could possible have different signals in their mRNA sequence leading to increased/decreased stability, or secondary structures. Furthermore, some or all of the identified variants might not be translated into protein, but can still have regulatory functions on the RNA level. This will need further elucidation. The identified transcripts with shorter 5' ends have interrupted signal and/or pro-region. One might speculate whether the interrupted signal sequence could lead to divergent transportation of the protein, and the non-existing pro-region to influence folding, stability and enzyme inactivation, but the possible consequences need to be further elucidated.

Finally, gene expression analysis of the different alternative transcripts in thymic tissue was performed. Since the disease associated haplotypes are present in the general population, gene expression could be measured in thymic samples from healthy individuals. It might, in fact, be advantageous to use tissue from healthy donors, since patients often receive immune-modulating treatment. All transcripts identified with the GeneRacer^{TM} method varied in the length of 5'- or 3' end, hence the expression level of these could only be measured indirectly by the assays utilized. In addition, because of the high sequence identity with cathepsin L (78%) (Santamaria et al. 1998), designing specific probes were further challenging. The assay covering the exon 5/6 boundary (assay 2) and the assay covering the exon 6/7 boundary (assay 3) did not detect higher expression than the one covering the exon 1/2 boundary (assay 1). This indicated that the shorter transcripts observed with GeneRacer^{TM} method, are not present in detectable higher amounts than full length *CTSL2* mRNA.

In order to relate gene expression with the associated haplotypes, the expression data was grouped according to the disease associated models shown by genotypic group analysis. A single SNP might have a crucial impact on a cell, but the combined effect of several polymorphisms in a haplotype might just as well contribute to disease development. It might be the exact constellation between alleles on a haplotype that together changes gene expression, mRNA folding, enzyme activity etc. This has earlier been shown for the *Apo E* gene, where polymorphisms in specific haplotype constellations influence the function of the protein (Fullerton et al. 2000). Nevertheless, no difference in *CTSL2* expression within the defined genotypic groups was observed. The relative expression of alternative variants might be so low that the differences are not detectable with the methods used, or 42 thymic samples might be too few to represent the variety present in the population. Another possible explanation might be that the disease models presented here are not the true models, or the polymorphisms not the aetiological variants. If further studies reveal that other polymorphisms in this gene area are involved, these analysis need to be readdressed. In addition, expression assays with specific probes for the two variants with part of intron 1 included should also be performed. No differences were found on the RNA level, but since regulation on protein level is just as important, further elucidation is warranted.

Our data suggest *CTSL2* as a susceptibility gene for development of T1D and MG, and an interesting candidate gene also for other autoimmune diseases. Knowledge of the interactions between polymorphisms in regulatory regions and disease predisposition will bring us further towards delineating why some individuals develop autoimmune diseases. Where medicine today treats after occurrence of disease, maybe the medicine of tomorrow can predict susceptibility before irreversible damage of tissue and organs occur. Thus, autoimmune diseases might be avoided, or age of onset delayed. To reach this goal, identification of the genetic variations leading to autoimmune diseases is fundamental.

Acknowledgement

We thank Christopher Bowlus for the fetal thymic cDNA sample and Harald Lindberg for the juvenile thymic tissue samples. This work was supported by the Juvenile Diabetes Foundation International, Novo Nordisk Foundation, Norwegian Diabetes Association, Research Counsil of Norway and the University of Oslo.

References

Barker DL, Hansen MS, Faruqi AF, Giannola D, Irsula OR, Lasken RS, Latterich M, Makarov V, Oliphant A, Pinter JH, Shen R, Sleptsova I, Ziehler W, Lai E (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. Genome Res 14:901-907

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265

Becker KG (1999) Comparative genetics of type 1 diabetes and autoimmune disease: common loci, common pathways? Diabetes 48:1353-1358

Becker KG, Simon RM, Bailey-Wilson JE, Freidlin B, Biddison WE, McFarland HF, Trent JM (1998) Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. Proc Natl Acad Sci U S A 95:9979-9984

Bennett ST, Lucassen AM, Gough SC, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, et al. (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9:284-292

Bottini N, Musumeci L, Alonso A, Rahmouni S, Nika K, Rostamkhani M, MacMurray J, Meloni GF, Lucarelli P, Pellecchia M, Eisenbarth GS, Comings D, Mustelin T (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat Genet 36:337-338

Caillat-Zucman S, Garchon HJ, Timsit J, Assan R, Boitard C, Djilali-Saiah I, Bougneres P, Bach JF (1992) Age-dependent HLA genetic heterogeneity of type 1 insulin-dependent diabetes mellitus. J Clin Invest 90:2242-2250

Concannon P, Erlich HA, Julier C, Morahan G, Nerup J, Pociot F, Todd JA, Rich SS (2005) Type 1 diabetes: evidence for susceptibility loci from four genome-wide linkage scans in 1,435 multiplex families. Diabetes 54:2995-3001

Dean J, Dean A, Burton A, Dicker R (1991) Statcalc Epi release 5.01b

Drachman DB (1994) Myasthenia gravis. N Engl J Med 330:1797-1810

Duff GW (2006) Evidence for genetic variation as a factor in maintaining health. Am J Clin Nutr 83:431S-435S

Engel AG (1984) Myasthenia gravis and myasthenic syndromes. Ann Neurol 16:519-534

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881-900

Furugaki K, Shirasawa S, Ishikawa N, Ito K, Ito K, Kubota S, Kuma K, Tamai H, Akamizu T, Hiratani H, Tanaka M, Sasazuki T (2004) Association of the T-cell regulatory gene CTLA4 with Graves' disease and autoimmune thyroid disease in the Japanese. J Hum Genet 49:166-168

Honey K, Benlagha K, Beers C, Forbush K, Teyton L, Kleijmeer MJ, Rudensky AY, Bendelac A (2002) Thymocyte expression of cathepsin L is essential for NKT cell development. Nat Immunol 3:1069-1074

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet 12:395-399

Huang D, Liu L, Noren K, Xia SQ, Trifunovic J, Pirskanen R, Lefvert AK (1998) Genetic association of Ctla-4 to myasthenia gravis with thymoma. J Neuroimmunol 88:192-198

Jacobson DL, Gange SJ, Rose NR, Graham NM (1997) Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol 84:223-243

Jaretzki A, III, Barohn RJ, Ernstoff RM, Kaminski HJ, Keesey JC, Penn AS, Sanders DB (2000) Myasthenia gravis: Recommendations for clinical research standards. Neurology 55:16-23

Johnson GC, Payne F, Nutland S, Stevens H, Tuomilehto-Wolf E, Tuomilehto J, Todd JA (2002) A comprehensive, statistically powered analysis of GAD2 in type 1 diabetes. Diabetes 51:2866-2870

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-1080

Kretz-Rommel A, Rubin RL (2000) Disruption of positive selection of thymocytes causes autoimmunity. Nat Med 6:298-305

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288:136-140

Maehr R, Mintern JD, Herman AE, Lennon-Dumenil AM, Mathis D, Benoist C, Ploegh HL (2005) Cathepsin L is essential for onset of autoimmune diabetes in NOD mice. J Clin Invest 115:2934-2943

Mehta C, Patel N, Gray R (1985) On computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables. Journal of the American Statistical Association: 969-973

Nakagawa T, Roth W, Wong P, Nelson A, Farr A, Deussing J, Villadangos JA, Ploegh H, Peters C, Rudensky AY (1998) Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus. Science 280:450-453

Nerup J, Pociot F (2001) A genomewide scan for type 1-diabetes susceptibility in Scandinavian families: identification of new loci with evidence of interactions. Am J Hum Genet 69:1301-1313

Nicholas K, Jr NH, Deerfield D (1997) GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEWS

O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259-266

Oosterhuis H (1997) Myasthenia Gravis. Groningen: Groningen Neurological Press

Pastinen T, Ge B, Gurd S, Gaudin T, Dore C, Lemire M, Lepage P, Harmsen E, Hudson TJ (2005) Mapping common regulatory variants to human haplotypes. Hum Mol Genet 14:3963-3971

Pastinen T, Ge B, Hudson TJ (2006) Influence of human genome polymorphism on gene expression. Hum Mol Genet 15:R9-R16

Pugliese A, Zeller M, Fernandez A, Jr., Zalcberg LJ, Bartlett RJ, Ricordi C, Pietropaolo M, Eisenbarth GS, Bennett ST, Patel DD (1997) The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. Nat Genet 15:293-297

Romi F, Gilhus NE, Aarli JA (2005) Myasthenia gravis: clinical, immunological, and therapeutic advances. Acta Neurol Scand 111:134-141

Santamaria I, Velasco G, Cazorla M, Fueyo A, Campo E, Lopez-Otin C (1998) Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. Cancer Res 58:1624-1630

Schneider S, Roessli D, Excofier L (2000) Arlequin: A software for population genetics data analysis release Ver 2.000

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-Wide Associations of Gene Expression Variation in Humans. PLoS Genet 1:e78

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487-498

Thorlacius S, Aarli JA, Riise T, Matre R, Johnsen HJ (1989) Associated disorders in myasthenia gravis: autoimmune diseases and their relation to thymectomy. Acta Neurol Scand 80:290-295

Tisch R, McDevitt H (1996) Insulin-Dependent Diabetes Mellitus Cell 85:291-297

Tolosa E, Li W, Yasuda Y, Wienhold W, Denzin LK, Lautwein A, Driessen C, Schnorrer P, Weber E, Stevanovic S, Kurek R, Melms A, Bromme D (2003) Cathepsin V is involved in the degradation of invariant chain in human thymus and is overexpressed in myasthenia gravis. J Clin Invest 112:517-526

Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, et al. (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 423:506-511

Undlien DE, Bennett ST, Todd JA, Akselsen HE, Ikaheimo I, Reijonen H, Knip M, Thorsby E, Ronningen KS (1995) Insulin gene region-encoded susceptibility to IDDM maps upstream of the insulin gene. Diabetes 44:620-625

Vafiadis P, Bennett ST, Todd JA, Nadeau J, Grabs R, Goodyer CG, Wickramasinghe S, Colle E, Polychronakos C (1997) Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. Nat Genet 15:289-292

Vandiedonck C, Capdevielle C, Giraud M, Krumeich S, Jais JP, Eymard B, Tranchant C, Gajdos P, Garchon HJ (2006) Association of the PTPN22*R620W polymorphism with autoimmune myasthenia gravis. Ann Neurol 59:404-407

Vincent A, Palace J, Hilton-Jones D (2001) Myasthenia gravis. Lancet 357:2122-2128

Wigginton JE, Abecasis GR (2005) PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. Bioinformatics 21:3445-3447

Supplementary

Supplementary table 1 Genotyping success rates and HWE tests for the investigated polymorphisms in the T1D family material and MG case-control material

Polymorphism	Position	Genotype success for T1D%	Genotype success for MG cases%	Genotype success for MG controls%	HWE in T1D	HWE in MG cases	HWE in MG controls
D9S1851	96650276-96650639	93	100	98	0.595	0.734	0.793
rs10739289	96874309	96	100	99	1.000	0.953	0.898
rs15394	96874680	87	100	98	0.986	0.953	0.809
rs7875800	96877315	97	100	99	0.660	0.953	0.846
rs4743056	96901166	96	100	99	0.981	0.605	0.999
D9S971	97246737-97246861	94	96	98	0.327	0.856	0.382

Positions of polymorphisms relative to the *CTSL2* gene; D9S1851: 223 kb downstream; rs10739289 (A/T): approx. 223 base pair downstream; rs15394 (A/C): in exon 8/3'untranslated region (UTR); rs7875800 (A/G): in the sixth intron; rs4743056 (T/C): 20 kb upstream; D9S971: 365 kb upstream. Marker positions were found at Ensembl (http://www.ensembl.org/index.html, v.36); HWE – Hardy-Weinberg equilibrium; HWE results are given as uncorrected p-values.

Supplementary table 2 Haplotype analysis of the *CTSL2* polymorphisms in the T1D trio material by the transmission disequilibrium test (TDT)

Haplotype	Alleles	T	NT	%T	95% CI	p-value
rs10739289-rs15394-rs7875800-D9S971	T-A-A-8	34	13	72.3	60-85	0.002*
rs10739289-rs7875800-D9S71	T-A-8	34	13	72.3	60-85	0.002*
rs10739289-D9S971	T-8	34	14	70.8	58-84	0.004
rs7875800-D9S971	A-8	34	13	72.3	60-85	0.002*
rs10739289-rs15394-rs7875800-rs4743056-D9S971	T-A-A-T-8	20	9	69.0	52-86	0.045
rs10739289-rs15394-rs7875800	T-A-A	80	57	58.4	50-67	0.060
rs10739289-rs7875800	T-A	83	57	59.3	51-67	0.040

T – number of transmitted alleles from heterozygous parents; NT – number of non-transmitted alleles from heterozygous parents; %T – percent transmission (deviation from 50% indicate association); 95% CI – 95% confidence interval for the %T. p-values are uncorrected; * corrected p-value=0.04

Supplementary table 3 Genotypic analysis of the associated two-SNP haplotype in the T1D material

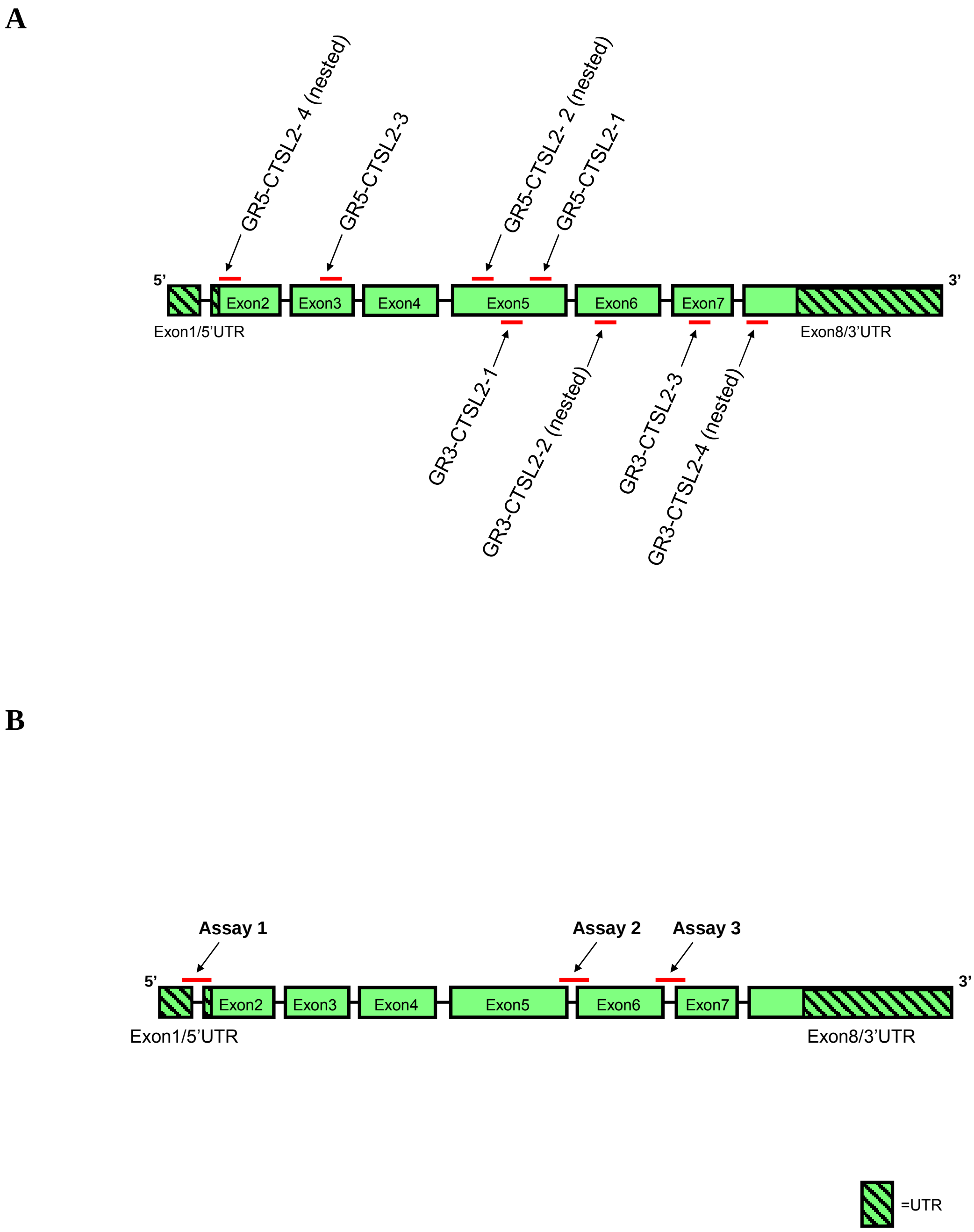
Haplotype groups	Cases n (%)	Family-based Controls n (%)	Statistical results
A/A	302 (83.0)	226 (75.8)	Chi^2 = 12.5
A/B	57 (15.6)	83 (22.8)	p=0.002[a] with 2 df
B/B	5 (1.4)	5 (1.4)	
A/A	302 (83.0)	226 (75.8)	OR for A/A vs. A/B and B/B = 1.9
A/B and B/B	62 (17.0)	88 (23.2)	95% CI = 1.29-2.79 , p=0.0006[b]

A: rs10739289-T and rs7875800-A alleles as a haplotype; B: rs10739289-A and rs7875800-G alleles as a haplotype; n – number of cases/controls; df – degrees of freedom; OR – odds ratio, CI – 95% confidence interval for OR, calculated with Cornfield 95% confidence limit; p-values are uncorrected; [a] – corrected p-value=0.04; [b] – corrected p-value=0.01

Supplementary table 4 Genotypic analysis of the associated haplotype in the MG material

Haplotype groups	All MG cases n (%)	EOMG cases n (%)	Controls n (%)
A/A	2 (2.5)	1 (3.3)	2 (0.8)
A/B	27 (33.8)	16 (53.3)	62 (25.8)
B/B	51 (63.8)	13 (43.3)	176 (73.3)
	Chi2 = 3.46 **p=0.18 with 2df**	**Chi2 = 11.91** **p=0.003[a] with 2 df**	
A	29 (36.2)	17 (56.7)	64 (26.7)
Non A	51 (63.8)	13 (43.3)	176 (73.3)
	OR for A vs. non A = 1.56 **95% CI = 0.87-2.76** **p=0.102**	**OR for A vs. non A = 3.6** **95% CI = 1.54-8.5** **p=0.0007[b]**	

A – rs4743056-C and D9S971-4; B – non MG haplotype; MG – myasthenia gravis; EOMG – early onset MG; n – number of cases/controls; df – degrees of freedom; OR – odds ratio; CI – 95% confidence interval for OR, calculated with Exact Confidence Limits (Mehta et al. 1985); p-values are uncorrected; [a] – corrected p-value=0.1; b – corrected p-value=0.03



Supplementary figure 1 Positions of primers and probes relative to the *CTSL2* gene. A) Positions of gene specific primers designed for GeneRacer^{TM}. GR3/GR5 – GeneRacer^{TM} 5'- or 3' end primer (example: GR3-CTSL2-1: GeneRacer 3' end-CTSL2-primer 1); B) Positions of probes for TaqMan® Gene Expression Assays. Green boxes indicate the full length (1369 bp) reference sequence of cathepsin L2 (*CTSL2*) mRNA from VEGA (http://vega.sanger.ac.uk); Primer and probe positions are marked red; UTR – untranslated region; all primer and probe sequences are given in supplementary table 5.

Primers and probes
Supplementary table 5

Experiment	Primer name	Sequence	Annealing	Dye Label	Quencher
Microsatellites					
	D9S971-F	5’-CCGCCACTCCTAAGGATG-3’	54 °C	FAM	
	D9S971-R	5’-CAGGTTGAACTATAAGCTCAC-3’	54 °C		
	D9S1851-F	5’-AAGTTCTATTCCCACAAAAAGAGAGT-3’	61 °C	FAM	
	D9S1851-R	5’-TGGCTTTGAGTTACTATGGTTCA-3’	61 °C		
Single nucleotide polymorphisms (SNP)					
TaqMan® primers					
	rs10739289-F	5’-GTAGGTTTGTCCTTCCAAATTGTTTATCAAATT-3’	60 °C		
	rs10739289-R	5’-GCTTCCTGGCTCTTGAGTAGTTATAAAAAT-3’	60 °C		
	rs4743056-F	5’-TCCCCAGTAGAAAAAAGTAAATGGCAA-T-3’	60 °C		
	rs4743056-R	5’-GGGAACTGTAACCACCCAAATTAAC-3’	60 °C		
	rs7875800-F	5’-GCAAAGCAGGATGTCATATCCAAGA-3’	60 °C		
	rs7875800-R	5’-CCAGTTCTACAAATCAGGTAAGTGTCATTTTATTATA-3’	60 °C		
	rs15394-F	Not available (Assay by demand)	60 °C		
	rs15394-R	Not available (Assay by demand)	60 °C		
TaqMan® probes					
	rs10739289-Reporter1	5’-AGTTACTGTGTGTTACCCAA-3’	60 °C	VIC	NFQ
	rs10739289-Reporter2	5’-TTACTGTGTGTAACCCAA-3’	60 °C	FAM	NFQ
	rs4743056-Reporter1	5’-TGATCATCACGCCTATG-3’	60 °C	VIC	NFQ
	rs4743056-Reporter2	5’-CTGATCATCACACCTATG-3’	60 °C	FAM	NFQ
	rs7875800-Reporter1	5’-ATCAATCACAGTGATGCT-3’	60 °C	VIC	NFQ
	rs7875800-Reporter2	5’-AATCACGGTGATGCT-3’	60 °C	FAM	NFQ
	rs15394-Reporter1	Not available (Assay be demand)	60 °C	VIC	NFQ
	rs15394-Reporter2	Not available (Assay by demand)	60 °C	FAM	NFQ
Context Sequence for rs15394:		5’-AATTAAAATCTCAACTTGGATCCTC[A/C]ATGATTCAACTGGTTTATCTTACAC-3’			
Splice variants					
	CTSL2-3’UTR-R	5’-TGAGTCTTTGATATCATAAAGCTGTG-3’	50-58 °C		
	CTSL2-Ex1A-F	5’-GCCGCCTGGAAACTTAAA-3’	50-58 °C		
	CTLS2-Ex1B-F	5’-TCTCAGAGGCTTGTTTGCTG-3’	50-58 °C		
	CTSL2-Ex2-F	5’-GTGGAAGGCAACACACAGAA-3’	50-58 °C		
	M13-F	5’-GTAAAACGACGGCCAG-3’			
	M13-R	5’-CAGGAAACAGCTATGAC-3’			

Experiment	Primer name	Sequence	Annealing	Dye Label	Quencher
GeneRacer^{TM} Gene Specific Primers					
	GR5-CTSL2-1	5’-AGGCCCTAGCCATGAAGCCACCATT-3’	78 °C		
	GR5-CTSL2-2	5’-CCTTCAAGAGCACCAGTCGCACTAA-3’	76 °C		
	GR5-CTSL2-3	5’-TGGCCATTGTGAAGCCATGTTTCC-3’	72 °C		
	GR5-CTSL2-4	5’-CCAGGACGAGCGAAAGATTCATGT-3’	72 °C		
	GR3-CTSL2-1	5’-ACTGTTCGCGTCCTCAAGGCAATCAG-3’	80 °C		
	GR3-CTSL2-2	5’-CACAGTGGTCGCACCTGGAAAGGAG-3’	80 °C		
	GR3-CTSL2-3	5’-GTTCTGGTGGTTGGCTACGGCTTT-3’	74 °C		
	GR3-CTSL2-4	5’-GGGGTCCAGAATGGGGCTCGAAT-3’	74 °C		
GeneRacer^{TM} Kit primers					
	GeneRacer^{TM} 5’ Primer	5’-CGACTGGAGCACGAGGACACTGA-3’	74 °C		
	GeneRacer^{TM} 5’-Nested primer	5’-GGACACTGACATGGACTGAAGGAGTA-3’	78 °C		
	GeneRacer^{TM} 3’ Primer	5’-GCTGTCAACGATACGCTACGTAACG-3’	76 °C		
	GeneRacer^{TM} 3’-Nested primer	5’- CGCTACGTAACGGCATGACAGTG-3’	72 °C		
	M13-F	5’-GTAAAACGACGGCCAG-3’	50 °C		
	M13-R	5’-CAGGAAACAGCTATGAC-3’	50 °C		
TaqMan® Gene Expression Assays: Primers					
	CTSL2EX1-2-M136-F	5’-AGGACGAGCGAAAGATTCATGTTT-3’	60 °C		
	CTSL2EX1-2-M136-R	5’-TGTAATCTCAGAGGCTTGTTTGCT-3’	60 °C		
	CTSL2EX5-6-M100-F	5’-AGCCAGTGTCATTAGCAACAGAATT-3’	60 °C		
	CTSL2EX5-6-M100-R	5’-CCTGGACTCTGAGGAATCCTATCC-3’	60 °C		
	CTSL2EX6-7-M95-F	5’-CTGCTGCAGTCTGGTTCAAAAT-3’	60 °C		
	CTSL2EX6-7-M95R	5’-ATGCAGGCCATTCGTCCTT-3’	60 °C		
Probes	CTSL2EX1-2-M136	5’-ACGGCTGCTGGTTTT-3’	60 °C	FAM	NFQ
	CTSL2EX5-6-M100	5’-CAGTGGATGAAATCTG-3’	60 °C	FAM	NFQ
	CTSL2EX6-7-M95	5’-AATGCCTGATTTGTAGAACTG-3’	60 °C	FAM	NFQ
	FG, HUMAN, B2M	Not available	60 °C	FAM	NFQ
	FG, 18S rRNA	Not available	60 °C	FAM	NFQ

NFQ: non fluorescent quencher