TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Medizin
Lehrstuhl für Humangenetik

Predicting the Efficiency of Interferon Therapy for Multiple Sclerosis using Genotype-based Machine Learning Models

Theresa Elena Schmiedlechner

Vollständiger Abdruck der von der

Fakultät für Medizin

der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Medizin (Dr. med.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Ernst J. Rummeny

Prüfer der Dissertation:

1. Prof. Dr. Bertram Müller-Myhsok

2. Prof. Dr. Johann Förstl

Die Dissertation wurde am 24.02.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 21.02.2018 angenommen.

Dissertation

Predicting the Efficiency of Interferon Therapy for Multiple Sclerosis using Genotype-based Machine Learning Models

Theresa Schmiedlechner

Thursday 9th February, 2017

dedicated to my family

Contents

List of Figures

List of Tables

1 Abstract

Despite extensive research, the pathogenesis of various autoimmune diseases still remains partly unresolved. For example, the cause of *multiple sclerosis (MS)*, one of the most common neurodegenerative autoimmune diseases, is still unknown and treatment approaches are limited. In most cases, interferon-β is an effective medication for MS Hartung et al. (2013); Sitzer and Steinmetz (2011). However, within time a large percentage of the patients treated with interferon-β produce binding antibodies (BABs) or neutralizing antibodies (NABs) which either bind or neutralize interferon-β and lead to therapy failure Creeke and Farrell (2013).

The aim of this thesis is to predict therapy response for interferon-β therapy by analyzing patients' genotypes. The data of MS patients treated with interferon-β as well as data on antibody development subsequent to medication and the genotype information were provided by the Neurological Department of Klinikum Rechts der Isar, Munich. We analyzed the data with a machine learning approach and discovered candidate genes that may be involved in antibody production in response to interferon-β treatment and might lead to a better understanding of the underlying molecular mechanism. So far the HLA-DRB1 gene and the SNP rs9272105, localized in close proximity to the HLA-DQA1 gene on chromosome 6, have been associated with antibody production against interferon-β Barbosa et al. (2006); Buck et al. (2011); Buck and Hemmer (2014); Hoffmann et al. (2008); Link et al. (2014); Soelberg Sorensen (2008); Weber et al. (2012). The SNPs rs4961252, localized on chromosome 8, and rs5743810, within the TLR6 gene on chromosome 4, also showed genome-wide significance, yet the latter was only the case in males whereas not in females Weber et al. (2012); Buck and Hemmer (2014); Enevold et al. (2010).

In this project, prediction models were created using machine learning techniques through the use of Support Vector Machines (SVMs). I wanted to go beyond single SNP effects and include SNP x SNP interactions in order to create a model based on candidate SNPs to predict a patient's response to medication for treatment of MS. Compared to other machine learning techniques, SVMs have the advantage of also accounting for SNP x SNP interactions.

In order to keep the number of SNP variants manageable for the SVM calculations, I partitioned the data in gene-wise subsets. For each gene-wise dataset, prediction models containing the SNPs that were ranked by their ability to predict antibody production were generated. These calculations resulted in a list of significant genes including the predictive features (SNPs). From these results I was able to identify the SNPs that achieved the best performance.

The results included HLA genes as well as the HCG23 and BTNL2 genes in close proximity on chromosome 6 to reveal significance. The SNP rs34784936, localized within the HLA region, achieved the best single SNP performance. Genome-wide, we found 78 genes with significant

results based on 315 SNPs. Of those, only the most relevant 166 SNPs need to be included in the final prediction model, since at that point the performance of the pruning calculation reaches its maximum. It is important to note that only a small set of selected genotype information of an individual patient is needed to predict therapy response. The identified genes associated with antibody production against interferon-β require further investigation.

2 Zusammenfassung

Trotz intensiver Forschung ist die Pathogenese verschiedener neurologischer Krankheiten bislang noch teils ungeklärt. So ist beispielsweise die Ätiologie der Multiplen Sklerose, einer der häufigsten neurodegenerativen Autoimmunkrankheiten, noch nicht vollständig bekannt und Therapieansätze sind nur eingeschränkt verfügbar. In den meisten Fällen stellt Interferon-β eine effektive Therapieoption dar Hartung et al. (2013); Sitzer and Steinmetz (2011). Dennoch entwickeln eine bedeutsame Anzahl der Patienten bindende Antikörper (BABs) oder neutralisierende Antikörper (NABs), die das Medikament binden bzw. neutralisieren und damit zu Therapieversagen führen Creeke and Farrell (2013).

Ziel dieser Arbeit war die Entwicklung eines auf Genotypen basierenden Vorhersagemodells, anhand dessen die Wahrscheinlichkeit der Antikörperbildung auf Interferon-β Medikation schon vor Therapiebeginn abgeschätzt werden kann. Darüber hinaus könnte man mögliche Kandidaten Gene identifizieren, anhand derer dann auf ein besseres Verständnis der molekularen Mechanismen gehofft werden kann, die dieser Krankheit und der Produktion von Antikörpern zugrunde liegen. Nach aktuellen Forschungsergebnissen liefern das Gen HLA-DRB1, sowie der SNP rs9272105, welcher in der Nähe des Genes HLA-DQA1 auf Chromosom 6 lokalisiert ist, erste Hinweise auf eine Assoziation von Antikörperproduktion als Reaktion auf eine Interferon-β Therapie Barbosa et al. (2006); Buck et al. (2011); Buck and Hemmer (2014); Hoffmann et al. (2008); Link et al. (2014); Soelberg Sorensen (2008); Weber et al. (2012). Auch die SNPs rs4961252 auf Chromosom 8 und rs5743810, welcher innerhalb des Gens TLR6 auf Chromosom 4 liegt, zeigten genomweite Signifikanz in Zusammenhang mit der Produktion von Antikörpern gegen Interferon-β letzterer jedoch nur bei männlichen Patienten Weber et al. (2012); Buck and Hemmer (2014); Enevold et al. (2010).

Mit der Fragestellung, ob anhand von genetischen Prädikatoren eine Vorhersage getroffen werden kann, wurden uns sowohl die Genotypen als auch die Daten zum Antikörpertiter gegen Interferon-β von der neurologischen Abteilung des Klinikums Rechts der Isar, München zur Verfügung gestellt.

Diese Dissertation beinhaltet Entwicklung eines Vorhersagemodels zur Antikörperproduktion gegen Interferon-β unter Berücksichtigung von SNP x SNP Interaktionen. *Support Vector Machines* ist eine Methode des maschinellen Lernens, die im Gegensatz zu anderen Methoden in der Lage ist, solche Interaktionen zu berücksichtigen. Dadurch geht dieses Modell über bisherige Forschungsansätze hinaus, die sich auf die Analyse von Einzel-SNP-Assoziationen oder maximal paarweisen Epistasiseffekten stützen.

Um die mögliche Anzahl der miteinbezogenen SNPs für eine SVM Berechnung nicht zu überschreiten, wurden die Genotypen genweise nach Gengrenzen aufgeteilt. Für jedes Gen

wurde ein Vorhersagemodel erstellt, das die zugeordneten SNPs entsprechend ihres Einflusses bezüglich einer Vorhersage zur Produktion von Antikörpern einstuft. Als Resultat ergab sich eine Liste signifikanter Gene mit den jeweils vorhersagerelevanten SNPs. Dadurch war es möglich, die vorhersagekräftigsten SNPs zu bestimmen.

Sowohl einige HLA Gene, aber auch die unmittelbar benachbarten Gene HCG23 und BTNL2 auf Chromosom 6 konnten als signifikant ermittelt werden. In den genomweiten Resultaten fanden sich 78 signifikante Gene mit 315 relevanten SNPs. Das endgültige Modell nutzt davon die 166 besten SNPs, welche die beste Vorhersage lieferten, da zu diesem Zeitpunkt bereits das Maximum der Vorhersage erreicht werden kann.

Wesentlich ist, dass für die zukünftige Anwendung dieses Modells nur ein ausgewählter Anteil der Genotypen eines Patienten zur Vorhersage benötigt wird. Dafür könnte man spezielle Tests entwickeln, die nur die im Modell verwendeten SNPs benötigen und somit relativ einfach und kostengünstig durchzuführen wären. Die identifizierten Gene sollten hinsichtlich ihrer Bedeutung weiter untersucht werden.

Part I

Introduction

3 Background

The specific cause of multiple sclerosis (MS) is still unknown, but it can partly be treated with interferon-β. The most commonly prescribed *Betaferon* is an immunomodulatory medication to prevent the occurence of acute relapses and nerve cell degeneration. Although this medication has a positive impact as far as reducing exacerbations and disease progress, a high percentage of patients produce antibodies against it. In this case, interferon-β is no longer recommended and other therapy arrangements must be considered Hartung et al. (2013); Weber et al. (2012). In order to avoid ineffective interferon-β treatment, constant examination and evaluation of the medication's activity needs to be performed. This way unproductive therapy can be detected and treatment can be adjusted for each patient individually. To improve this situation, a method to predict therapy failure beforehand is desired.

Unfortunately, patients can not yet be identified on the basis of clinical data whether they are at risk of developing binding or neutralizing antibodies or if they will respond well to therapy. Being able to predict therapy response through the analysis of selected biomarkers of an individual's genome indicates a promising improvement in future medicine.

A recent study points out that the discovery of predictive biomarkers is of great interest in ongoing multiple sclerosis research Buck and Hemmer (2014). So far the genetic markers primarily localized in HLA regions on chromosome 6 have been associated with antibody production against interferon-β. In particular these are the HLA-DRB1 gene and the SNP rs9272105 localized in close proximity to the HLA-DQA1 gene Barbosa et al. (2006); Buck et al. (2011); Buck and Hemmer (2014); Hoffmann et al. (2008); Link et al. (2014); Soelberg Sorensen (2008); Weber et al. (2012). Furthermore, the SNP rs4961252 localized on chromosome 8 showed genome-wide significance Weber et al. (2012); Buck and Hemmer (2014). Also the SNP rs5743810, within the TLR6 gene on chromosome 4 revealed a correlation to the production of antibodies against interferon-β in males, whereas not in females Enevold et al. (2010). The discoveries of these possible genetic risk factors influencing the antibody production against interferon-β motivated us to start our project—the aim of creating a model based on the genotype data to predict therapy response for patients on interferon-β medication. Recently, more and more studies focus on DNA-analysis, the investigation of the function of genes, and their coded proteins, or on single SNP examinations, which can lead to changes of the phenotype when mutated. With the prospect of knowing specific allele-disease associations, individual genetic predisposition may be recognized even before disease outbreak. In consequence studies, analyzing the genome with regard to possible association to a disease (e. g., *GWAS*) increasingly gain in interest not only for research but also to the general public. Various companies (e. g., 23andMe) offer a genome-wide marker analysis to find out more about individual carrier

status, health risk (genetic predisposition) and drug response. This leads to an increasing number of DNA examinations and consequently to a larger data pool of genotype information. The German company, STADA Diagnostik analyses their patients' genotypes to improve the predictive power in treatment response. The program performs laboratory tests of the genome for individual and optimal therapy stategy and consequentely supports the attending physician on his decision which substance is the most suitable for an individual patient.

Although there are some known genetic associations of antibody production in response to interferon-β therapy, the single SNP effects are too weak to yield reasonable prediction power. Therefore, in our project we searched for a method considering interactions. This way, a prediction power beyond the single SNP effects can be achieved. Furthermore, we aimed for a method that is able to detect indicative SNPs associated with antibody production against interferon-β, which raise prediction power and may indicate possible new biomarkers. In this thesis we used Support Vector Machines to create a prediction model.

Support Vector Machines (SVM) is a machine learning technique which has been employed successfully in classification as well as in regression analysis Bennett and Campbell (2000); Cantor-Rivera et al. (2014); Toshimoto et al. (2014). Genotype information of multiple sclerosis patients treated with interferon-β, as well as the phenotype corresponding to the antibody titer against medication, provided us all the data needed.

The ability to predict how well interferon-β medication will be tolerated by a patient affected with multiple sclerosis would mean a major leap forward in treatment procedures. Knowing a patient's risk to develop antibodies beforehand would be a better way to avoid ineffective medication. Furthermore, by adjusting and optimizing medication early, a reduction of treatment time and costs can be achieved. The goal of this project is to develop a prediction model for which only a small amount of genotype information of an individual patient is needed and which ultimately can be obtained readily in the future.

4 Multiple Sclerosis

4.1 Introduction

Multiple sclerosis is a chronic autoimmune inflammatory disease of the central nervous system with a mean global prevalence of 33 per 100 000 Multiple Sclerosis International Federation (2013). The autoimmune reaction is of unknown cause and leads to axonal impairment and demyelination of nerve cells in the brain and spinal cord. This causes a steady decrease of brain function. Women are affected twice as often as men, mostly with disease onset between 20 and 40 years of age. Although the etiology of MS is not yet understood, various risk factors such as viral infections (e. g., Epstein-Barr virus), nicotine, vitamin D deficiency, or genetic predisposition are discussed. The prevalence of multiple sclerosis is in fact higher in Europe, northern America and Australia than around the equator and regions with warm and tropical climates, as shown on a global map in figure 4.1.

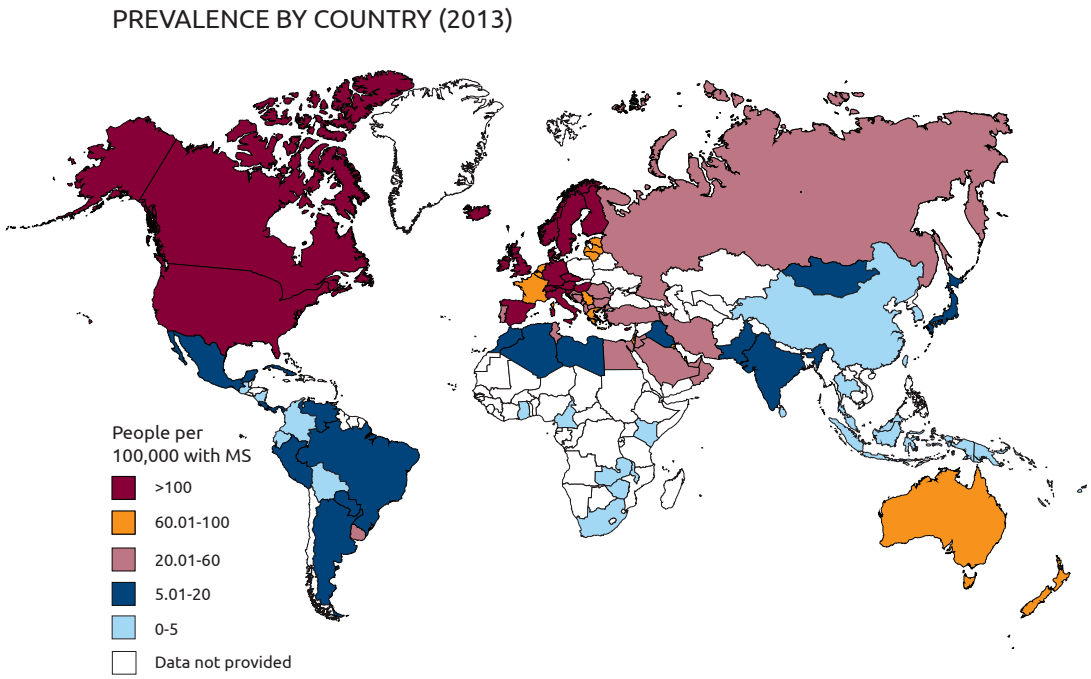

Figure 4.1: Prevalence of multiple sclerosis. Figure retrieved from the *Atlas of MS 2013* by the *Multiple Sclerosis International Federation*, 2013 available for download at www.msif.org/wp-content/uploads/2014/09/Atlas-of-MS.pdf, Multiple Sclerosis International Federation (2013).

Genetic heredity is also a noticable factor in risk for multiple sclerosis. In recent years, various genetic markers have been identified in association with the disease — specifically genes that are known to be responsible for the expression of immunomodulatory agents influencing the immune response. In particular, allele variants of the HLA-DRB1 gene on chromosome 6 are correlated to predisposition of MS Baranzini (2011); The International Multiple Sclerosis Genetics Consortium and the Wellcome Trust Case Control Consortium 2 (2011); Sitzer and Steinmetz (2011). Additionally, genes located in other regions of the genome could be detected recently Baranzini (2011); The International Multiple Sclerosis Genetics Consortium and the Wellcome Trust Case Control Consortium 2 (2011). Fig. 4.2 on the facing page shows an overview of genomewide potentially associated regions beyond the major histocompatibility complex region.

Genetic biomarkers have not only been associated with disease outbreak and progress, but have also shown correlation to therapy response. See section 4.2.4 for details, more examples will be described.

4.1.1 Symptoms

The first person to describe characteristics and pathology of multiple sclerosis in detail was the French neurologist Jean-Martin Charcot, who in 1868 defined this clinical picture as *sclérose en plaques disseminées* Hafler (2004). He defined *staccato speech*, *nystagmus*, and *intention tremor*, also known as the *'Charcot's triad'*, to be the three characteristic symptoms for MS Sitzer and Steinmetz (2011).

Today, a broader spectrum of symptoms is considered. In early manifestation, the most common symptoms reported in MS are:

- sensory disturbance such as numbness or tingling in fingers,
- unilateral optic neuritis resulting in blurred or double vision, and
- lack of coordination.

During the course of the disease, the following can be affected:

- the motor system, resulting in weakness or paresis
- the sensory system, resulting in numbness or tingeling, paraesthesia, or pain
- the sense of vision, resulting in reduced visual acuity
- the brainstem, resulting in cranial nerve disorders (e. g., trigeminal neuralgia, facial nerve paresis, nystagmus)
- the cerebellum, resulting in lack of coordination, intentional tremor and ataxia
- the vegetative system, resulting in bowel or bladder disturbance
- and in advanced stages of disease, even the cognitive function of the brain can be affected, resulting in attention deficit and reduced memory performance.

In short, any symptom can emerge depending on where the inflammatory focus is localized in the brain.

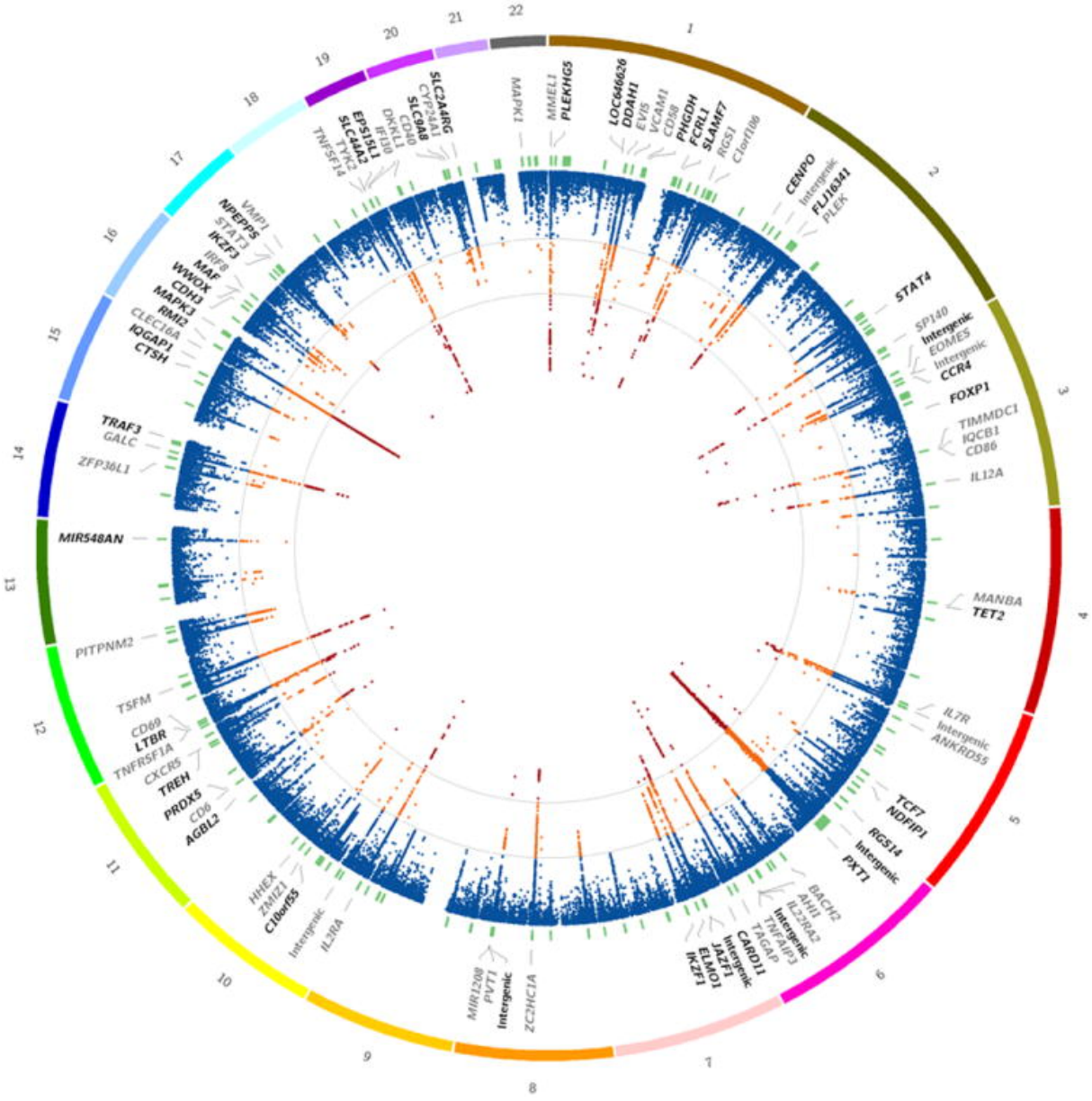

Figure 4.2: "Circos plot showing primary association analysis of 161,311 autosomal variants in the discovery phase (14,498 cases and 24,091 healthy controls). The outermost track shows the numbered autosomal chromosomes. The second track indicates the gene closest to the most associated SNP meeting all replication criteria. Previously identified associations are indicated in gray. The third track indicates the physical position of the 184 fine-mapping intervals (green). The innermost track indicates $-\log(P)$ (two-sided) for each SNP scaled from 0–12, which truncates the signal in several regions" caption citation and image retrieved from the published study *Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis* by the *International Multiple Sclerosis Genetics Consortium (IMSGC)* in *Nature Genetics* (2013) available at www.nature.com/ng/journal/v45/n11/pdf/ng.2770.pdf, International Multiple Sclerosis Genetics Consortium (IMSGC) et al. (2013).

4.1.2 Forms

80–90 % of MS patients are affected primarily by the *relapsing remitting* form of multiple sclerosis (RRMS) Hartung et al. (2013); Sitzer and Steinmetz (2011). This form is characterized by fulminate attacks, also called exacerbations of autoreactive activated $CD4^+$ T-cells on the myelin of nerve cells. The symptoms must persist longer than 24 hours and be separated by a minimum of 30 days from the last incident to be considered an exacerbation. Furthermore, these cells produce cytokines and recruit even more immunocompetent cells such as macrophages, B-cells, and natural killer T-cells, which leads to an augmentation of the inflammatory process and eventually results in neuronal loss and gliosis.

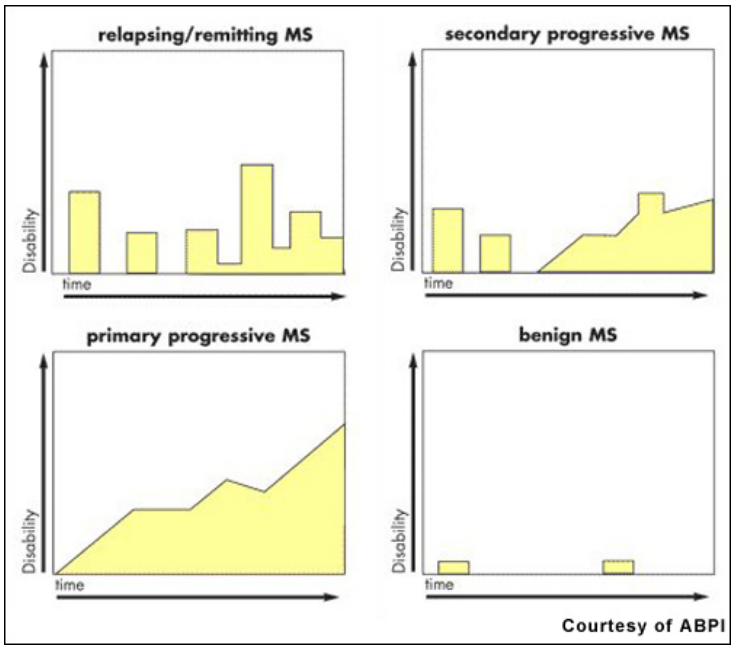

Figure 4.3: MS types and course of disease. Image retrieved from "Fingolimod - Novel Therapy for Multiple Sclerosis" by the *Association of British Pharmaceutical Industry (ABPI)* available at www.drugdevelopment-technology.com/projects/fingolimod/fingolimod1.html, Association of British Pharmaceutical Industry, ABPI (ny).

The occurence of attacks causes a sudden worsening of persisting symptoms or results in the onset of a new symptom. Yet, the RRMS is characterized by the partial regression of symptoms within 6–8 weeks. In some cases the symptoms can even fully dissolve in case of regression. Without treatment, a large percentage of RRMS evolve into *secondary progressive* multiple sclerosis (SPMS) within ten years. This form of MS is characterized by a continuous (but not necessarily rapid) degradation and function loss with less frequent exacerbations. 10–15 % of all MS cases are diagnosed with *primary progressive* multiple sclerosis (PPMS) Hartung et al. (2013); Sitzer and Steinmetz (2011). Patients with PPMS experience disease progression from the very beginning. Their conditions then worsen without years of remission. They are more likely to experience problems with walking and steady progression of symptoms whereas in relapsing attacks, sudden worsening is more frequent. The *progressive relapsing* form (PRMS) is the least common form of MS. It is a combination of RRMS and PPMS, appearing in initial progression and may be accompanied by occasional relapses. Some patients are also affected by a very mild and *benign* form of the disease, which is characterized by rare incidents of mild severity. The figure 4.3 shows an overview of the different forms of MS.

4.1.3 Diagnosis

The initial demyelinating incident or inflammatory episode in the central nervous system is referred to as a *clinically isolated syndrome* (CIS), and may go on to develop into multiple sclerosis.

To diagnose MS there has to be neuroradiological evidence of at least two separate lesions in the white matter of the brain, as well as their occurence at different points in time. Some examples of typical MRI lesions of multiple sclerosis are shown in Fig. 4.5 on the next page, appearing periventricular, subcortical, infratentorial including the spinal cord, and around the corpus callosum, also known as *'Dawson's fingers'*, see Fig. 4.4.

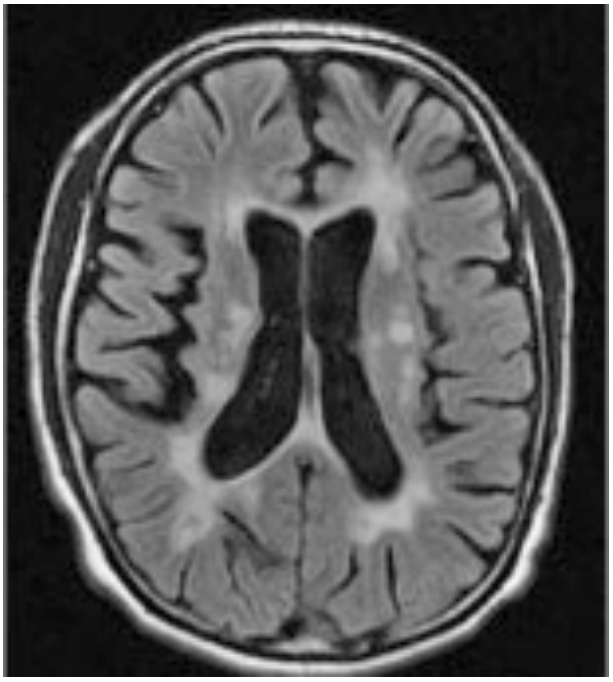

Figure 4.4: Dawson's fingers. Image retrieved from the picture archive of the magnetic resonance imaging of the Max-Planck-Institute of Psychiatry, Munich.

Apart from MR imaging, the analysis of the *cerebrospinal fluid* (CSF) can indicate a possible evolving multiple sclerosis — represented as oligoclonal bands, intrathecal immunoglobulin G (Ig-G) synthesis and a mild pleocytosis in the CSF.

Due to the demyelination in the nervous system, the nerve cells show a reduced nerve conduction velocity. This means they can not transmit the action potential and deliver information as fast as healthy cells. The nerve cell capacity can be measured by *visual* or *acoustic evoked potentials* (VEP/AEP). A high latency and decreased amplitude of the recorded potentials — a typical symptome within this disease — indicate a pathologic demyelinating process.

4.1.4 Progression and prognosis

Multiple sclerosis can appear and proceed in various different clinical presentations within each individual patient. So far, there are no reliable criteria or parameters to predict an individual's course of disease. Nevertheless, risk factors may reflect the severity of affection. Factors such as

- young age at diagnosis,
- female sex,
- the RRMS form with a small number of exacerbations and full recovery, as well as
- predominantly only sensory symptoms in early manifestation

are known to be beneficial for a mild progression.

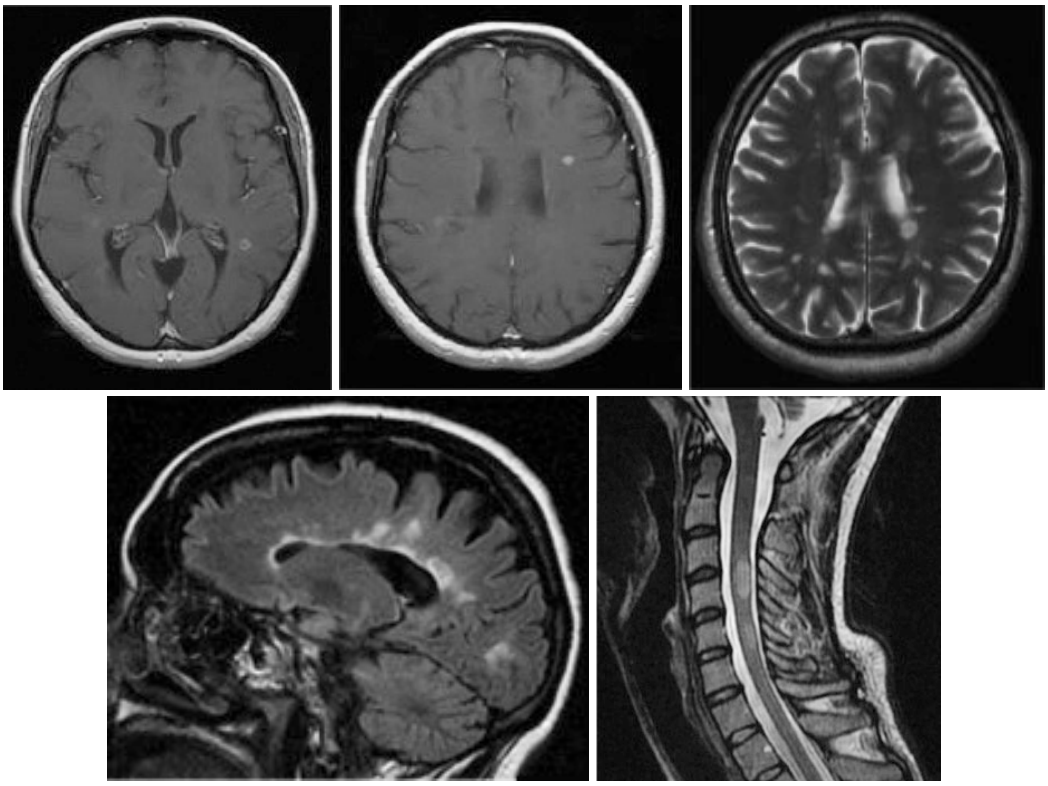

Figure 4.5: MR images of the brain and spinal cord showing typical MS lesions. On the top row demyelinating lesions are presented in T_1-, gadolinium enhanced T_1- as well as T_2-weighted sequences in transversal plane. Bottom row: lesions especially around the corpus callosum (left) and in the spinal cord (right) are shown in a *Fluid Attenuated Inversion Recovery* (FLAIR) and T_2-weighted MR image. Images retrieved from the picture archive of the magnetic resonance imaging of the Max-Planck-Institute of Psychiatry, Munich.

In comparison, the following indicate unfavorable fast progression:

- a high age at diagnosis,
- male sex,
- the PPMS form or a high number of exacerbations with polytope and cerebellar symptoms, or
- pyramidal tract impairment.

To classify a patient's condition based on clinical examination, the *Expanded Disability Status Scale* (EDSS-Score) has been established. The EDSS-Score includes evaluation of the current function of the brain (including the brainstem, cerebellum, and vision), motor and sensory function, as well as bowel and bladder regulation. The EDSS ranges from 0, indicating a completely normal neurological examination, to 10, death due to multiple sclerosis. Up to an EDSS score of 3 a patient is unrestrictedly able to walk. Depending on the manageable walking distance and further restrictions the score rises as shown in detail in Table 4.1 on the next page.

EDSS	clinical performance
0	normal neurological examination
1	no disability, minimal abnormalities
2	minor disability
3	moderate disability, no restriction of walking
4	severe disability, able to walk 500 meters unassistedly
5	severe disability, able to walk 200 meters unassistedly
6	walking aid necessary to walk a distance of 100 meters
7	incapable of walking a distance of more than 5 meters, restricted to wheelchair
8	restricted to bed, able to use arms
9	restricted to bed, able to communicate and eat
10	death due to disease

Table 4.1: Expanded Disability Status Scale, adapted from Grehl and Reinhardt (2013), p. 468.

4.2 Treatment guidelines for multiple sclerosis

Although multiple sclerosis can not yet be cured, there are medications to treat acute attacks, to improve the patients general state of health, and options for long-term medication. These options will be presented in the following section.

4.2.1 Therapy for acute relapses

In the event of an acute attack, *corticosteroids* is an efficient standard form of therapy. High doses of methylprednisolone (500–1000 mg) are applied intravenously for 3 to 5 days. In addition, plasmapheresis may be considered.

4.2.2 Long-term therapy

Long-term therapy is recommended in early stages of disease to prevent the occurrence of acute relapses as well as to protect functional nerve cells from degenerating. The most common medication is *interferon-β*, which will be discussed in the next section (4.2.3). *Betaferon*, a subcutaneously (s. c.) applied interferon-β_{1b} is considered a gold standard medication for MS treatment Hartung et al. (2013). In case of intolerance or counter indications to interferon-β, *Glatirameracetat* (Copaxone®) is a good alternative. Glatirameracetat is a combination of various amino acids, which imitate myelin to recover the nerve sheath. During the past few years new effective medications have been approved for severe or therapy-resistant cases of MS, which need to be prescribed carefully with regard to each individual medical situation. *Natalizumab* is a recombinant monoclonal antibody to detain circulating leukocytes from passing the blood brain barrier. Thus, some cases of progressive multifocal leukoencephalopathy have been reported after monthly infusions of Natalizumab. *Mitoxantron* is an immunosuppressive and cytostatic drug. Yet, due to its high cardiotoxicity blood levels and cardial function need to be reviewed before application. *Dimethylfumarat*, *Teriflunomid*, *Alemtuzumab*, or *Fingolimod* may provide other promising therapy alternatives.

4.2.3 Interferon-β therapy

Interferon-β is one of the most prescribed medications for MS patients and is a promising attempt to delay disease progress and reduce exacerbations. It was approved in the USA in 1993 for MS therapy and is also counted as the first registered medication for RRMS Hartung et al. (2013). Although its impact on the immune system is not yet fully known, it is assumed interferon-β has

- antiviral,
- antiproliferative (suppression of pro-inflammatory cytokines and increased production of anti-inflammatory agents), and
- immunomodulatory (inhibition of T-cell proliferation and increased apoptosis as well as down-regulation of MHC-II expression)

effects, which achieve remarkable therapy success. In case of RRMS, interferon-β can reduce the number of exacerbations by over 30 % after two years of therapy. There is a 32 % reduction when treated with Avonex, 33 % with Rebif and even up to 34 % less exacerbations when treated with Betaferon compared to placebo prescription Hartung et al. (2013). Furthermore, interferon-β lowers the occurence of T_2 active lesions in an MRI, and prevents the appearance of new lesions. This may be explained by the neuroprotective and regenerative impact of interferon-β on neurons.

All together, interferon-β can not only slow down disease progress from CIS to clinically manifest MS, but also delay the development from RRMS to SPMS.

After many years of clinical trials on this medication's effect, it can be prescribed in case of RRMS and SPMS, as well as in early stages of disease, CIS, or for patients at high risk to develop MS. Interferon-β is a protein of 165 amino acids and can be categorized in two subtypes:

Interferon-β_{1a}	is obtained from mammalian cells
Interferon-β_{1b}	is obtained from *E. coli* bacteria or synthetically produced

Table 4.2: Subtypes of Interferon-β Hartung et al. (2013).

Interferon-β can by applicated subcutaneously (s. c.) or intramuscularly (i. m.), depending on dose and medication as listed in Table 4.3.

Interferon	Subtype	Application		Dose
Betaferon	Interferon-β_{1b}	s. c.	250 µg	every other day
Avonex	Interferon-β_{1a}	i. m.	30 µg	weekly
Rebif 22	Interferon-β_{1a}	s. c.	22 µg	3× per week
Rebif 44	Interferon-β_{1a}	s. c.	44 µg	3× per week

Table 4.3: Interferon-β medication for multiple sclerosis treatment Hartung et al. (2013).

In case of newly emerging attacks during interferon-β therapy, the dose can be increased. Although it is a commonly prescribed medication and generally well tolerated, we should not forget the frequently occuring side effects, such as:

- influenza-like symptoms like fever, fatigue, and shivering,
- skin irritations or infiltrations and necrosis at injection site,
- headaches,
- alterations of the blood count as, e. g., anaemia, leukopenia and thrombocytopenia as well as lymphopenia,
- alterations of the liver function caused by drug elimination process in liver and kidney,
- myalgia, or
- mood swings.

4.2.4 Antibodies against interferon-β

Treatment with interferon-β can induce the production of binding or neutralizing antibodies. These antibodies bind the applied interferon-β and may inhibit its inpact on human cells consequently leading to therapy failure. Binding antibodies may be produced very early after the beginning of interferon-β therapy. Although their occurence does not necessarily lead to treatment failure, they may indicate a larger chance of producing neutralizing antibodies later in time Creeke and Farrell (2013). A study reveals that in up to 45 % of multiple sclerosis patients treated with interferon-β a production of neurtralizing antibodies can be observed Creeke and Farrell (2013). This occurs mostly after a time period of 6 to 18 months of therapy, which is even more often observed with s. c. than i. m. application Hartung et al. (2013). To avoid ineffective medication, it is necessary to frequently evaluate interferon activity. To detect interferon-β activity *in vivo* the *myxovirus resistance protein (MxA)* gene expression is the most commonly measured parameter. During interferon-β medication, an increased transcription of MxA mRNA can be observed. In case of antibody production, neutralizing antibodies bind interferon-β and attenuate its effect. This leads to lower concentrations of MxA which can, therefore, be considered a reliable variable to clinically measure therapy efficiency. Better yet, to avoid ineffective medication, is it is best to try to recognize beforehand if the therapy with interferon-β has a chance of being successful. In other words, genetic or clinical biomarkers, which are used to predict treatment response, are needed. So far, genetic markers primarily localized in HLA regions on chromosome 6 have been associated with antibody production against interferon-β. In particular these are the HLA-DRB1 gene and the SNP rs9272105 localized in close proximity to the HLA-DQA1 gene Barbosa et al. (2006); Buck et al. (2011); Buck and Hemmer (2014); Hoffmann et al. (2008); Link et al. (2014); Soelberg Sorensen (2008); Weber et al. (2012). Furthermore, the SNP rs4961252 localized on chromosome 8 showed genome-wide significance Weber et al. (2012); Buck and Hemmer (2014). Also the SNP rs5743810, within the TLR6 gene on chromosome 4 revealed a correlation to the production of antibodies against interferon-β in males, whereas not in females Enevold et al. (2010). In case antibodies are detected in a patient's blood serum, current treatment guidelines recommend to consider therapy rearrangements Hartung et al. (2013); Weber et al. (2012).

5 Exploring Support Vector Machines (SVM)

5.1 Background

Support Vector Machines (SVM) were first introduced by Corinna Cortes & Vladimir Vapnik in 1995 as a machine learning approach for two-group classification and yet represent a popular technique of *kernel methods* used for classification and regression analysis Cortes and Vapnik (1995). Kernel methods like *SVM* and *kernel principal component analysis* are machine learning techniques and used to recognize patterns such as rankings, correlation, principal components, classification or regression in high-dimensional data Schölkopf et al. (1997). SVMs in particular have shown to be successful in various applications Bennett and Campbell (2000); Cantor-Rivera et al. (2014); Toshimoto et al. (2014).

SVM can easily be applied where information of high-throughput technologies needs to be filtered to extract the relevant subset of parameters to answer specific medical questions. Applications can be envisioned for a classification into e. g., disease subtypes, responder/non-responder cases, or even more specifically, for prediction of treatment outcome.

The concept behind and the application of SVMs will be discussed in the following sections using classification and regression analysis, the kernel methods and implementation of SVM will be described.

5.2 Classification

In classification problems, one aims to find an optimal separation between two or more classes based on some measured parameters. This type of problem is found whenever the outcome is described by a categorial variable, which can indicate for example a disease, given medication or a patients country of origin.

SVM can perform automated classification of unknown cases based on their specific combination of the measured parameters.

5.2.1 Separable classes

A simple example is presented in Fig. 5.1 on the next page, where two classes, represented by the different symbols, are classified into two groups. In the two dimensional case shown here, SVM determines the optimal separation simply by a straight line. In three dimensions, we would

require a separating plane to classify data. In general, the separating feature in n dimensions is a separating subspace of dimension $n - 1$, which in higher-dimensional space is referred to as hyperplane. The *optimal* hyperplane is defined as having the largest possible distance between the closest data points of opposite classes. Those points which are located right on the boundaries are called *support vectors*. The simple illustration in Fig. 5.1 will help visualize this concept. Examplary displayed are data points from two classes in the upper left (indicated by the ○ symbols) and lower right (◇), respectively, depicting a simple two-dimensional linearly separable classification problem.

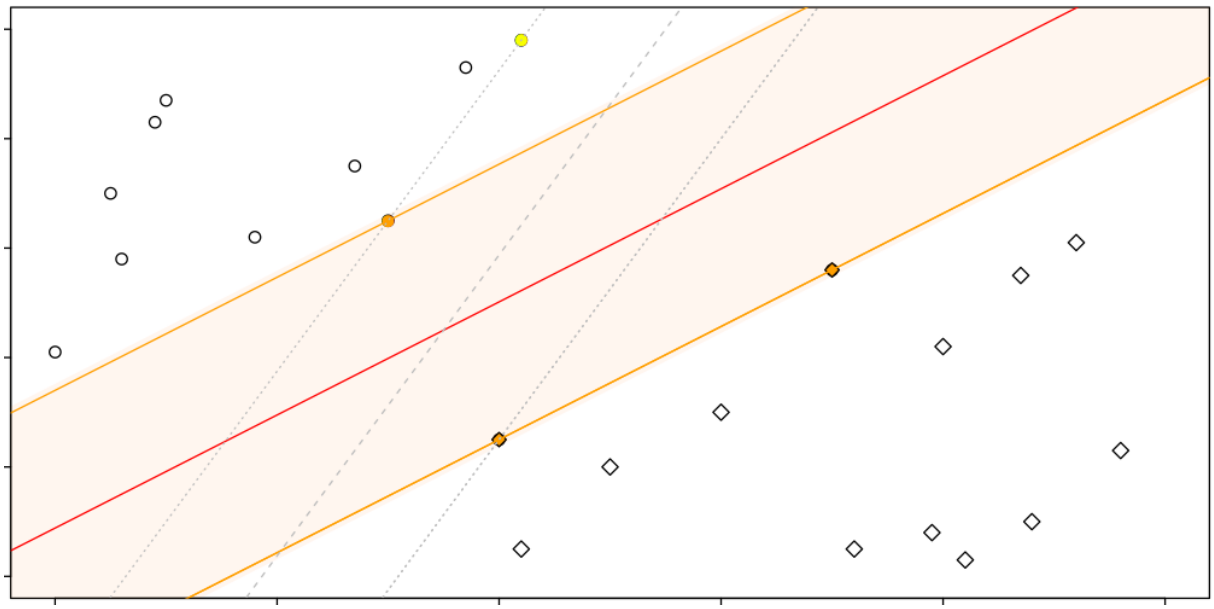

Figure 5.1: SVM classification of two classes (○ and ◇ symbols). The red line indicated the best possible separation. The margin is represented by the orange coloured area, the orange coloured circles indicate the support vectors. The dashed grey line demonstrates a non-ideal separation.

While there are actually many possible separating lines, the red line indicates the one with the *largest possible* distance between the nearest points of the different classes. The so-called *margin* is indicated by the orange colored area. As long as data points do not lie in the orange area, they do not contribute to the separation. The filled orange circles are called the support vectors as they — and only they — define this optimal separation line. The more complex data is arranged , the more support vectors are needed to define the separating hyperplane. We included another optional dividing hyperplane, the dashed grey line, to demonstrate a non-ideal separation, as the corresponding margin, indicated by the dotted grey lines, is clearly more narrow than the orange one, which is not what one is aiming for.

5.2.2 Non-separable classes

With real data, the classes may not be linearly separable at all. Under such circumstances, no matter how the hyperplane is placed, one or more data points will lie on the wrong side of the separating hyperplane, and, therefore, be classified incorrectly.

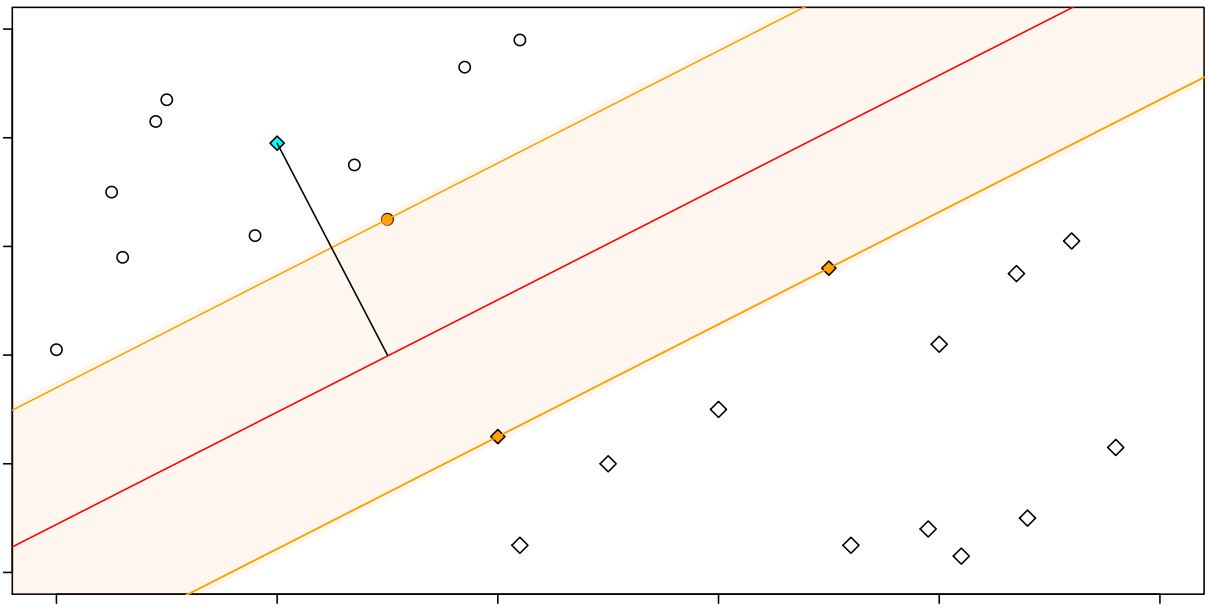

Figure 5.2: SVM classification showing a blue colored data point on the wrong side of the hyperplane.

Penalties In order to still find an ideal separation, a penalty is introduced for each misclassification, usually depending on the distance of the data point to the hyperplane. The blue colored outlier from the ◇ class in the lower right and its distance to the red separation line are illustrated in Fig. 5.2. The hyperplane for which the minimum total penalty ensues is considered *optimal*. As this margin no longer provides a clear separation, it is called a *soft margin*. In this case all the data points on the boundaries as well as those on the wrong side, which means either classified correctly or incorrectly, are support vectors.

Data transformation In some cases, not even the soft margin allows an acceptable solution and an alternative way to handle non-separable classes may be to use a different representation of the data. A transformation to a different coordinate system may help separate the classes more easily. An example is shown in Fig. 5.3 on the following page, where the original Cartesian coordinate system is defined by the measured variables, here denoted x and y. The separation, indicated by the dotted circle, does not allow a linear separation at all. However, if the data are represented in polar coordinates the situation depicted in the right plot results which is clearly separable.

SVMs can perform such kind of transformations effortlessly through appropriate use of kernels, see section 5.4 on page 35.

5.3 Regression

When working with continuous data, we use SVMs to create a regression model. Compared to the classification approach, the aim is now to fit all the data points *within* the margin. This will set the regression line close to as many values as possible. Illustrated in Fig. 5.4 on

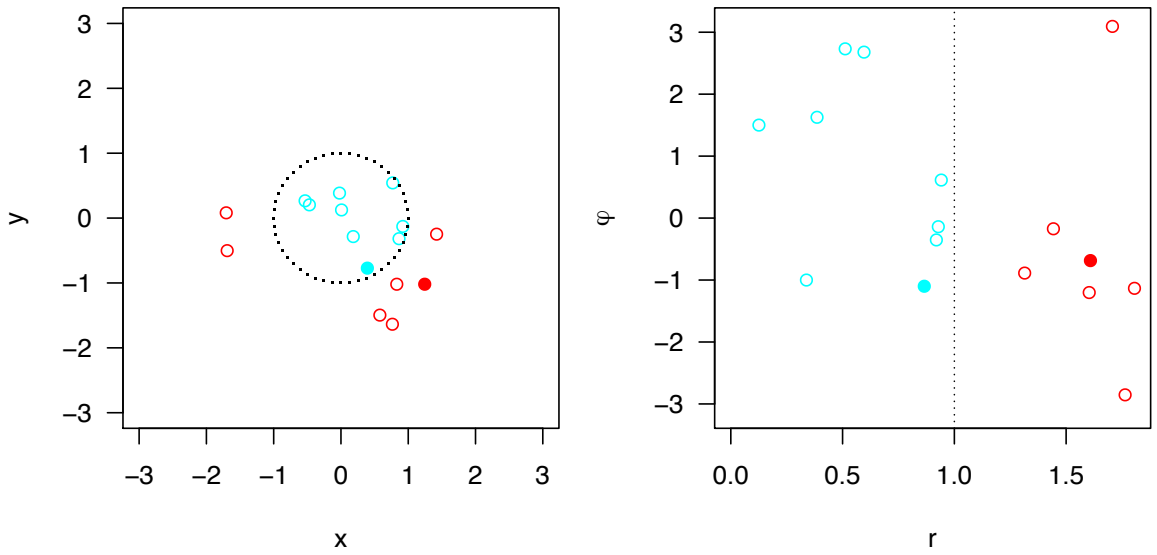

Figure 5.3: Transformation to a more appropriate coordinate system may result in a separable representation of the data. In the example shown, the circular distribution on the left is transformed to a linear one (right) by going from Cartesian to polar coordinates.

the following page are the regression line in red, half way between the support vectors as orange colored circles, and the orange shaded area enclosing as many data points as possible. Nevertheless, when working with real data, it is not always possible to place the regression line ideally adjusted for all values. In consequence, some values have to be penalized as in the non-separable classification case, illustrated by the white circles outside the orange area. Again, this way outlier effects can be minimized.

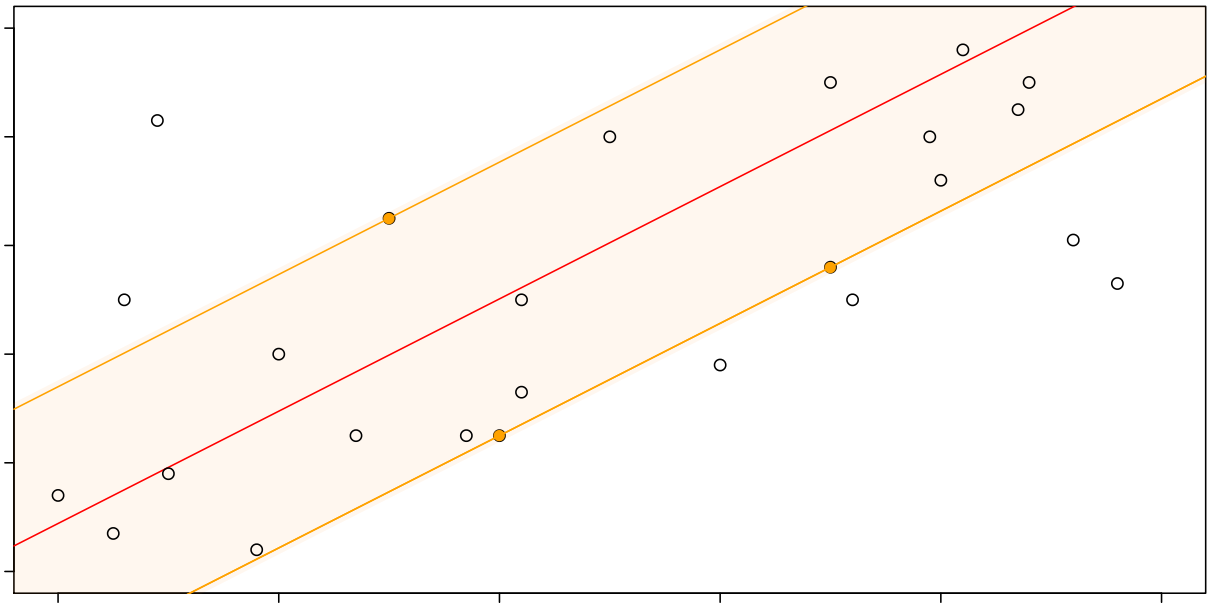

Figure 5.4: SVM regression. The red line illustrates the regression line, the orange coloured area represents the margin and the orange coloured circles indicate the support vectors.

5.4 Kernel trick

When working with real data, it may not always be possible to find a proper separation of the data using hyperplanes even in low dimensional feature spaces. Additionally, as the dimensionality or complexity of the data increases, it becomes progressively difficult to find a suitable separation. The so-called *kernel trick* resolves this problem by transforming the raw data into higher dimensional space in such a way that the transformed data becomes separable and the SVM approach can be applied. Through utilization of the kernel trick, the data points and the resulting separating hyperplane are only represented through *dot products* and the transformations can be calculated by the kernel functions, listed in Table 5.1 on page 38. This means the following:

Within the raw data each feature represents one dimension in the input space. For example, a dataset of 50 SNPs localized on one gene defines a 50-dimensional feature space. Every individual is represented by a 50-element vector indicating the corresponding SNP's genotypes.

The *kernel trick* virtually transforms the data into high dimensional space in such a way that data becomes separable. The data and the separating hyperplane are now only represented through *dot products*.

The larger a dataset gets, the more difficult it is to comprehend and follow the computations. Compared to the transformation to the polar coordinate system, the kernel trick can easily calculate from *only the coordinates in the original feature space*. The *kernel trick* only *virtually* transforms data in high dimensional space. Therefore, the important aspect of this procedure is that the calculations are performed in the low-dimensional input space. This is possible since the virtually-created separating hyperplane and support vectors in high dimensional space can be transformed back into the original space. This is shown for example in Fig. 5.5 on the following page retrieved from the *DTREG - predictive modeling software* website, which shows the complex separation in the original input space obtained from a separation performed in virtually contructed high dimensional space. This means, by making use of the *kernel trick*, there is no need to actually calculate the transformation to obtain the *dot products* in the new coordinates. All calculations can be performed in the original input space, which makes calculations practical and computationally feasible. This is a great advantage compared to the transformation to polar coordinates, introduced in section 5.2.2 on page 32, where this form of computation was not possible and transformed data were used for further calculations.

5.5 SVM prediction models

As previously explained, SVM creates classification or regression models in dependence of a particular feature or characteristic of the data, e. g., disease, given medication, age at diagnosis, and many others. In this study the phenotype indicates the antibody titer against interferon-β medication, as further introduced in chapter 6. This means for this particular study that a prediction of the antibody titer can be calculated with the given genotype data of an individual treated with interferon-β. The details of these calculations are explained in the following:

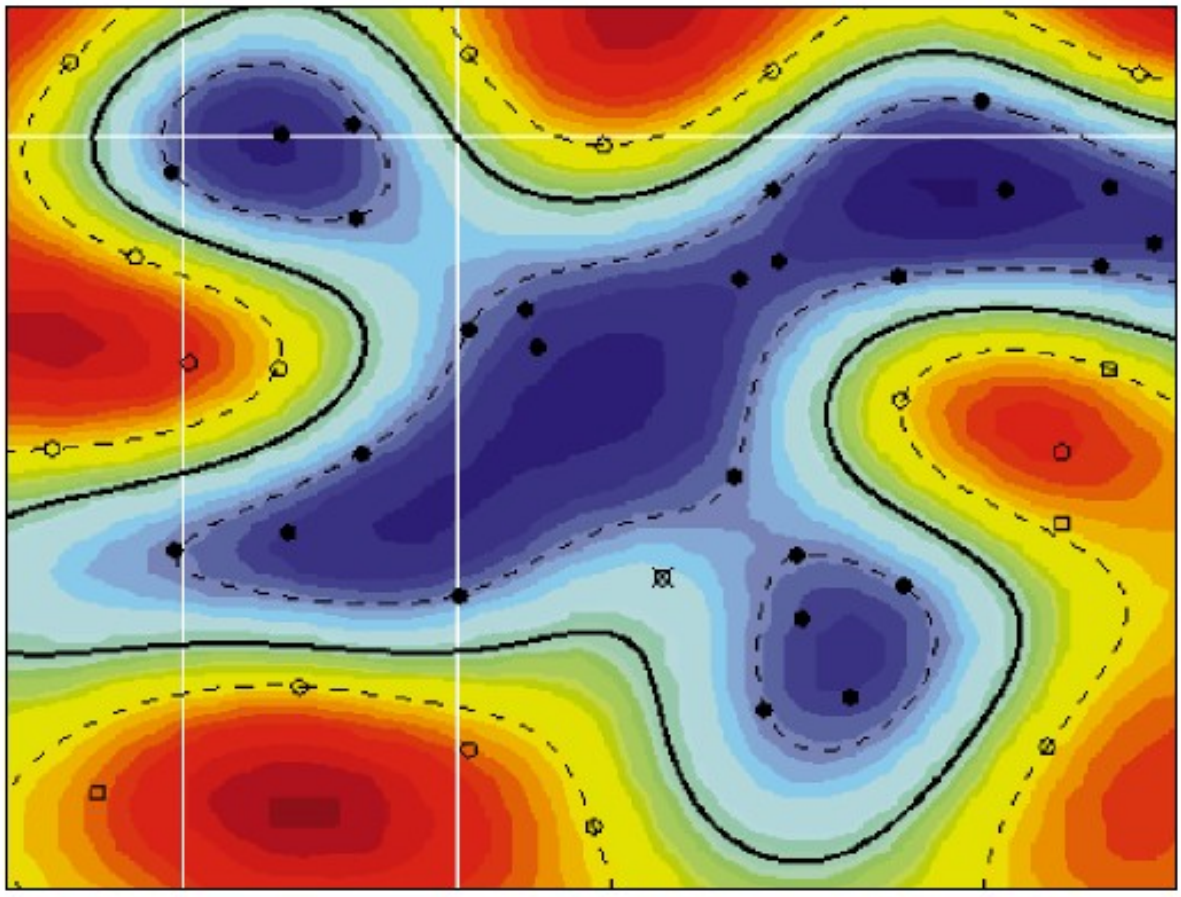

Figure 5.5: Non-linear separation of data in the original input space by the kernel trick. Figure retrievd from the *DTREG - predictive modeling software* website illustrating Support Vector Machines (SVMs), available at www.dtreg.com/solution/view/20, DTREG - predictive modeling software (2014).

First, data is separated into a trainingset and testset. Then SVM creates a prediction model on the basis of the training set data. To assess SVM model efficiency and to evaluate the accuracy and performance of SVM, the generated model can be reviewed on the testdata. The constructed model is therefore applied to examine and analyze a subset or testset of the data. For each individual of the testdata a predicted antibody titer is generated. With the correlation coefficient between the measured and predicted values, model performance can be calculated.

Correlation coefficient The correlation coeffictient, denoted by the r-value, is defined as the extent of the similarity of two variables. The r-value ranges from -1, indicating negative, over 0 showing no, to 1, indicating absolut correlation. A difference between the Pearson correlation, which describes linear coherence, and the Spearman correlation, which demonstrates the monotonic coherence of rank transformed data can be seen. Their outcome can differ from each other, but data containing no correlation will result in r-values close to zero.

The square of the correlation coefficient, r^2, describes the total variance of the data. This means that by calculating r^2 we find how much of the variance of the dependent variable can be explained by the influence of the independent variable. For example, an r^2 of 0.80 indicates that 80 % of the variance of the observation can be explained by the influencing variable. For our study we are interested in the correlation between measured and predicted antibody titer to interferon-β. A high correlation of the measured and predicted antibody titers indicates a high validity of the prediction model. For more detail on the implementation of SVM, see the next section 5.6.

Although SVMs yield advantages to other machine learning approaches such as, e. g., the ability to directly find interactions, computing reproducible results and many others, they are not perfect. Known shortcomings of SVM are that calculations with a lot of data tend to overfit. This means they create overly good models when working on too many parameters. Regardless that in this case SVM can forcast prediction values almonst perfectly with this data, it cannot necessarily be implemented reliably for another dataset. In such cases, some preprocessing (excluding uninteresting, or highly correlated parameters and splitting data into suitable partitions) is required to avoid this problem as will be explained in detail in the next chapter 7.2.2.1 on page 69.

5.6 Implementation of SVMs

The standard software base for SVM implementations is *libsvm 2.6*, developed by Chih-Chung Chang and Chih-Jen Lin Chang and Lin (2011).

Interfaces to many programming languages are available. We use the software **R** R Core Team (2014), for which various packages with SVM implementations are available, general ones such as, e. g., **e1071** Meyer (2012); Meyer et al. (2015), **kernlab** Karatzoglou et al. (2004, 2016), or more specific ones like **penalizedSVM** Becker et al. (2009, 2012) for feature selection in classification problems.

After initial test runs had shown the packages almost identical in their outcome, we decided on using the **R**-package **e1071** for our calculations.

The syntax of the `svm` call is as follows:

```
svm.model <- svm(formula, data, cost, gamma,
  type, cross,  kernel, ...)
```

The parameters to the `svm()` command are selected and the calculation results are assigned to variable `svm.model`.

`formula` indicates the dependent and independent features for which the model should be created, meaning in this study the antibody titer against interferon-β should be predicted.

`data` matrix containing phenotype (dependent feature) and genotype (independent) information.

`cost` is a penalizing SVM parameter for wrong classification as explained in section 5.2.2. It has to be regarded when data points appear on the incorrectly classified side of the hyperplane. This means that in case of inseparable data penalties must be considered in the classification model, depending on the distance of the misclassified data point to the hyperplane.

`gamma` γ is a kernel parameter specific for the Gaussian kernel, which determines the reach of a features influence. See Table 5.1 and as further explained in section 7.2.2.3 on page 72.

`type` indicates the form of classification or regression, e. g., *C-classification*, *ν-classification*, or *ε-regression*, ..., where `svm` can automatically choose between classification and regression depending on the type of the dependent variable.

`cross` determines the sampling method to be used. If `cross` $= n$ is specified, an n-fold cross validation will be performed. This means when using the example of `cross` $= n$, the dataset will be divided into n partitions. The training set data is used to create the prediction model. The model is subsequently used to examine and analyze the remaining data, referred to as testset data, to assess model efficiency. The maximum value allowed for `cross` is N, the number of individuals in the data, resulting in dividing the data into N parts, evaluating N models with one individual removed. This case is also called *leave-one-out cross validation (LOOCV)*, where, as the name indicates, all except one individual is used as training data. Without specification of `cross`, all data will be included for the model creation. According to our test results, using various n range from 3 to 5 over 100 or even n, we did not observe differences, so we employed $n = 3$ for performance reasons.

`kernel` is by default set to *radial basis kernel*, also called *Gaussian kernel*, used when having normally distributed data as it is the case with our data. Other kernel types readily available for `svm` within the **R**-package **e1071** are listed in Table 5.1 with their respective parameters Meyer et al. (2015).

Kernel	formula
linear	$u'v$
polynomial	$(\gamma u'v + \text{coef0})^{\text{degree}}$
radial basis	$e^{(-\gamma\lvert u-v\rvert^2)}$
sigmoid	$\tanh(\gamma u'v + \text{coef0})$

Table 5.1: List of commonly used kernel functions, the respective *kernel parameters* are set in green. Adapted from the arguments documentation for `svm` within the reference manual of the **R**-package **e1071** Meyer et al. (2015).

The function `svm()` returns a list of components, which summarize model features, method and results. This includes a summary of parameter values, such as `cost` and `gamma` and also `type`, `cross` or `kernel` chosen for the SVM model. Among others, `SV`, the number of support vectors, is returned. This gives an indication of how complex the separation of the data

points needed to be. A high number of support vectors indicates difficult separation, whereas a low number is usually found for easily separable data.

In the following step, the prediction on the test set is performed,

```
predicted.values <- predict(svm.model)
```

which yields the predicted values of antibody titer for each individual. To estimate the performance of the predictive model, some measure of concordance is needed.

We use the correlation coefficient between the measured and predicted values, which can be calculated, to evaluate model performance.

```
r <- cor(predicted.values, measured.values)
```

Higher values of r indicate better prediction.

To visualize prediction outcome, we plot the measured values against the resulting predicted antibody titer values for each data points

```
plot(predicted.values, measured.values)
```

as shown in figure 5.6. This is an example of the HLA-DRB1 gene, with gene boundaries extended by ± 10 kb. On the top of the figure you can see that the data contains 103 features (101 SNPs as well as the covariates sex and age) of 354 individuals. In this calculation, 326 support vectors were needed to compile the optimal regression line. An $r - value$ of 0.428 could be reached.

Since a more detailed discussion of the mathematical formulation of SVMs lies beyond the scope of this thesis, the interested reader may find more extensive explanation in the standard literature see, e. g., *Introduction to Statistical Learning with* ***R*** James et al. (2014) or *Elements of Statistical Learning* Hastie et al. (2009).

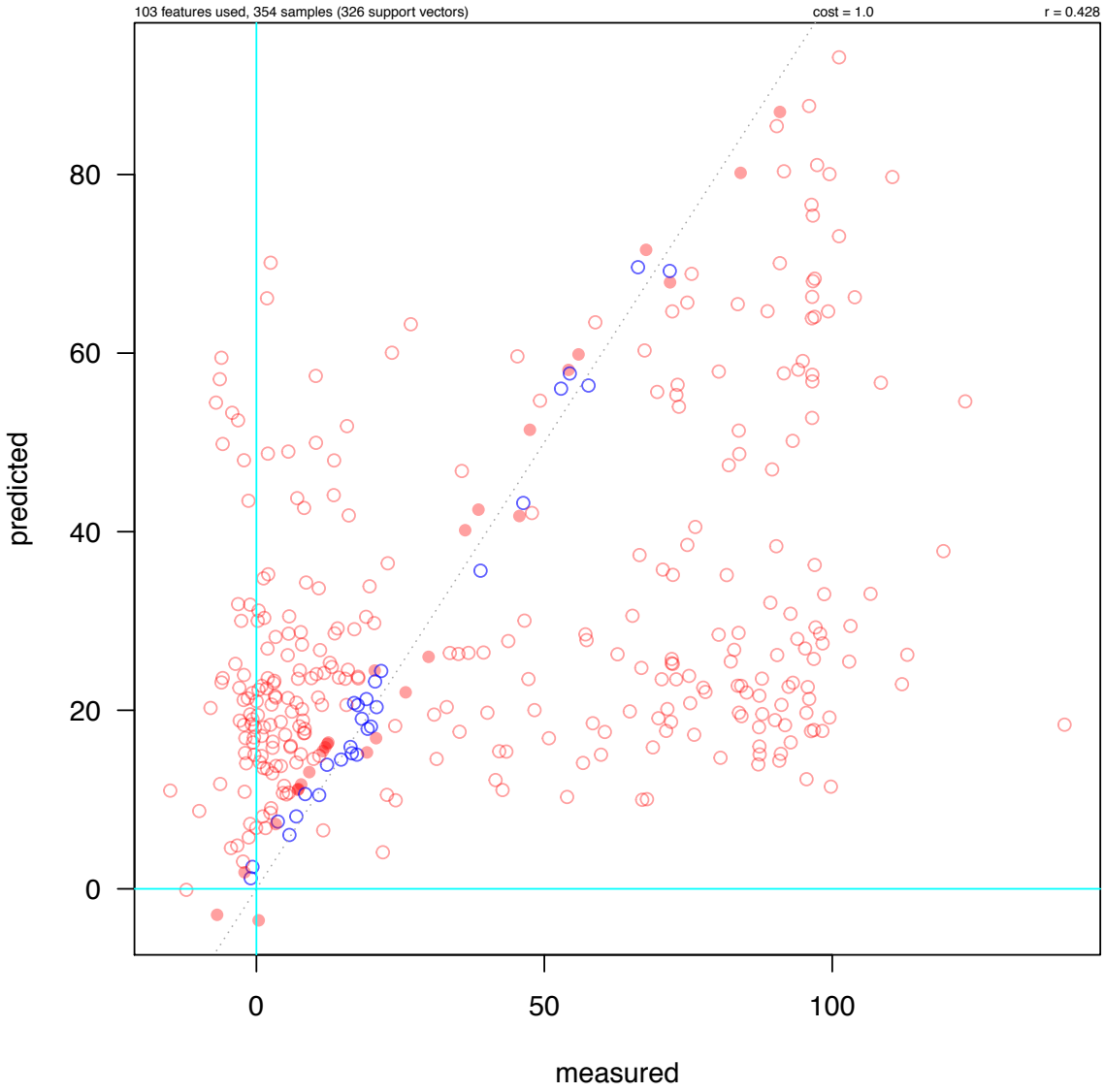

Figure 5.6: Prediction plot of the HLA-DRB1 gene. The dataset contains 101 SNPs (and the covariates sex and age) of 354 individuals. The regression line and support vectors are shown. An r-value of 0.428 could be reached.

Part II

Methods

6 Data Preparation

Within this project we worked with two different datasets containing genotype and phenotype features of multiple sclerosis patients treated with interferon-β. For initial calculations, including the creation of the SVM model as well as within the working process of constant reevaluation, we used a dataset consisting of 392 individulas. For better unterstanding, this dataset will be referred to *TUM 1 dataset* throughout my thesis and will be presented in detail in section 6.1. While working on my project a larger dataset was created by the Department of Neurology at the Rechts der Isar Hospital, affiliated to the Technical University of Munich. It displays a larger dataset of 1000 individuals, partly overlapping with the *TUM 1 dataset* and will be introduced as the *TUM 2 dataset* in section 6.2 of this thesis. To achieve an even larger sample size, we merged the two datasets to a combined study, *TUM 3 dataset*, which was used for final calculations, evaluation and interpretation of results and will be presented in section 6.3.

6.1 TUM 1 Dataset

6.1.1 Raw Data

For model preparations, genotype data of 392 multiple sclerosis patients was used. As explained in the published article, *Single-nucleotide polymorphism in HLA- and non-HLA genes associated with the development of antibodies to interferon-β therapy in multiple sclerosis patients* by Weber et al. in *the Pharmacogenomics Journal* (2012), where the same data was used, the genotyping was performed as explained in the following citation: "Genome-wide genotyping was performed by HumanCNV 370-Duo_v1-0 BeadChip (Illumina, San Diego, CA, USA) arrays, which covered about 317.000 single-nucleotide polymorphism (SNP) loci from the entire human genome. Genotyping was performed according to the standard protocols of the manufacturer for the Infinium II process" Weber et al. (2012).

The dataset is composed of 229 female and 125 male MS patients, aged between 16 and 75 years at the time of sampling. Besides their genomic sequence, for 354 patients a list of phentotype characteristics were recorded by the Neurological Department at the Rechts der Isar Hospital in Munich, Germany, which include features and covariates such as sex, age, disease progress, EDSS, medication, antibody titer against interferon-β, and many others. In detail, for each individual information on their medication status such as start, duration, and efficiency was recorded. The majority of the patients (169) were treated with *Betaferon*, another 134 patients were treated with *Rebif 44*, an interferon-β_{1a}. *Rebif 22*, also an interferon-β_{1a} in lower

dosis, was prescribed to 35 patients. 16 patients received *Avonex*, the only intramuscularly (i. m.) applied interferon-β_{1a}, as listed in Table 6.2.

The possibility of therapy failure due to antibody production against interferon-β is why it is extremely important to continually assess each patient's antibody status. Any antibody production against interferon-β can be detected through *enzyme-linked immunosorbent assay (ELISA)*. This method allows the examiner to obtain the antibody titer against interferon-β of each patient, however not distinguish between *binding antibodies* (BABs) and *neutralizing antibodies* (NABs). For this reason, the MxA concentration is also measured, which indicates the antibody reaction to interferon-β and therefore reveals the mediacation's residual function. In this study, an antibody reactivity of at least 25 % (100 % indicating the highest positive control, 0 % no antibodies) was considered *antibody-positive*. Patients with high measured antibody reactivity were either classified to NABs cases when an MxA induction of less than 50 % was observed or BABs when MxA concentrations exceeded $50\,\%$. Patients developing binding antibodies may still show some interferon-β activity, although reduced. No antibody reactivity or titer values below 25 % where counted as *antibody-negative* status, independent of the MxA induction, as shown in the overview in Table 6.1.

Antibody status	Antibody titer	MxA Induction
Antibody positiv	$\geq 25\,\%$	
∟ Neutralizing antibodies		$< 50\,\%$
∟ Binding antibodies		$\geq 50\,\%$
Antibody negativ	$< 25\,\%$	

Table 6.1: Antibody titer classification.

This classification splits the sample into 172 antibody positive and 182 antibody negative cases, the positive cases further subdivided into 45 binding and 127 neutralizing antibody cases. A detailed listing of antibody status with respect to medication is compiled in Table 6.2.

Medication	Patients	AB negative		BABs		NABs	
Betaferon	169	87	51.5 %	26	15.4 %	56	33.1 %
Avonex	16	6	37.5 %	3	18.8 %	7	43.8 %
Rebif 22	35	19	54.3 %	3	8.6 %	13	37.1 %
Rebif 44	134	70	52.2 %	13	9.7 %	51	38.1 %
Total	354	182		45		127	

Table 6.2: Antibody status by interferon-β medication.

Of all patients, only those individuals with a measured antibody titer in the extreme of the distribution were selected for genome sequencing and included in the study. This explains the two marginal density peaks seen in the distribution on the left side of Fig. 6.1 on the next page. To regain approximately normally distributed data, as required for some statistical methodes, the antibody titer values were replaced by their rank position using the *inverse rank-based transformation*, which is a favoured normalization technique when working with non-normally distributed data. Its result, the *normalized* antibody titer, is shown on the right side of Fig. 6.1.

The subsequent procedures are performed with both the original and the normalized data.

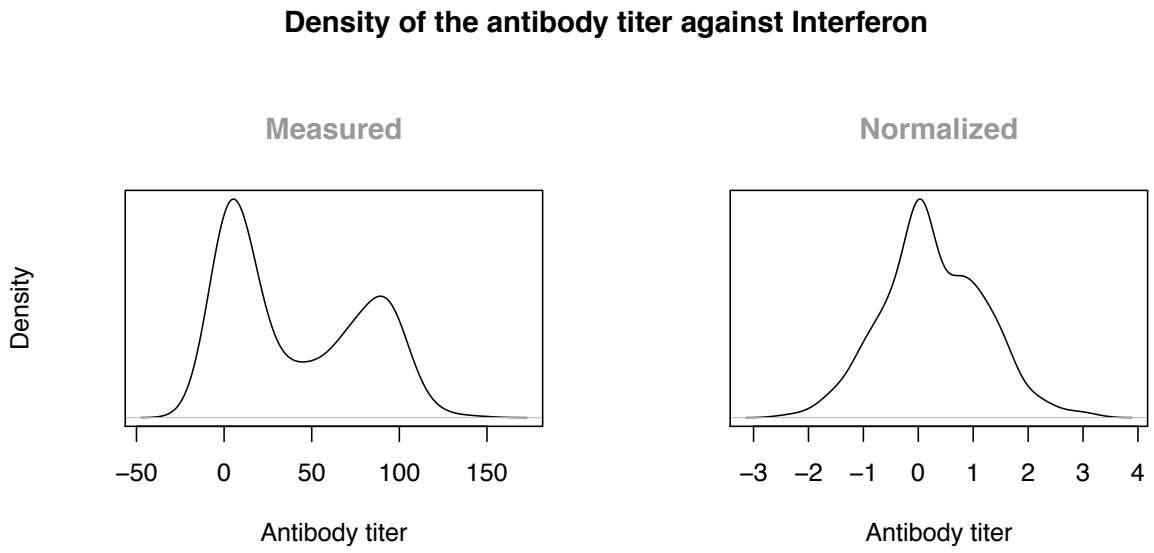

Figure 6.1: Density plot of the antibody titers of the 354 individuals included in the study. The left figure shows the distribution of measured antibody titers. The right figure shows the normalized values after *inverse rank-based transformation* to regain approximately normally distributed data.

Within this study we did not differentiate the measured antibody titers due to its effect on interferon-β. For following calculations we utilized the antibody titers detected with *ELISA*, which again means no conclusion on therapy outcome can be made.

6.1.2 Quality Control

In order to obtain reliable results from genotypic data, an initial quality control analysis of the dataset needs to be performed. This ensures that the data is free from obvious errors or inconsistencies, which may severely affect study outcome. It is a crucial procedure comprising data inspection, filtering, and examination.

Depending on the data format, various steps have to be completed. We received data imputed with SHAPE IT, which is used for pre-phasing, together with IMPUTE2, a genotype imputation program Howie et al. (2009, 2011); Delaneau et al. (2012, 2013). Imputation is a method to complete missing genotype data by implementing and predicting absent data using the known genotype structure of the sample and information from reference data sets, e. g., HAPMAP or 1000genomes, with more complete genotype maps The International HapMap Consortium (2003); The 1000 Genomes Project Consortium (2012, 2015). HAPMAP includes about 4 million SNPs exemplary for a specific population, 1000genomes even some 40 million SNPs. This means that in our case with genotype data, missing allele information- such as *single nucleotide polymorphism* (SNP) - can be predicted based on the knowledge of genotype relations of the reference population data. The raw genotype data was prephased using SHAPE

IT and subsequentely imputed using Impute version 2.

A *single nucleotide polymorphism* (SNP) is a variation of one single base pair in the DNA sequence, e. g.,

Allele 1	···	C	A	**T**	G	T	···
Allele 2	···	C	A	**G**	G	T	···

In this case, there are two alleles: **T** and **G**. They describe the variability of genotype data at a specific gene locus in a population. In some cases, this base exchange can lead to conformational changes in the transcribed protein which can manifest in a clinical phenotype, such as specific characteristics or even disease development. It is likely that highly correlated SNPs within so-called *linkage disequilibrium* (LD) blocks will be found.

The knowledge of these correlations makes a prediction possible. LD blocks reflect the non-random correlation of alleles. This means that alleles or genetic markers occur more or less often than expected by chance in a specific combination, a haplotype. In other words, LD is the deviation from random. LD relates the major (p_A and p_B) or minor (p_a and p_b) allele frequencies of the individual genotypes to their combined occurrence.

When the genotypes of two SNPs are (statistically) independent, the expected probability of an allele combination in a haplotype is given by the product of the individual allele probabilities, i.e.,

$$p_{AB} = p_A \cdot p_B. \tag{6.1}$$

Genetically, this can be interpreted as a recombination or mutation and referred as *linkage equilibrium*.

In contrast, *linkage disequilibrium* (LD) describes the situation where we find Eq. 6.1 violated, i.e.,

$$p_{AB} \neq p_A \cdot p_B. \tag{6.2}$$

The degree of violation is given by the disequilibrium coefficient D_{AB} which can be calculated as the difference of observed and expected probability:

$$D_{AB} = p_{AB} - p_A \cdot p_B \tag{6.3}$$

Due to the known association of alleles, either within the sample or from more extended populations, missing genotypes can be predicted. Furthermore, only one SNP in an LD block needs to be kept in a study as the remaining genotypes in the block are not providing independent information and can be left out, as realized within `preFilter()` in 7.1.2.2 to reduce sample size.

Genotype data is represented by the *minor allele frequency (MAF)*, which is the occurrence of the least common (= minor) allele in a population. Rare allele variants, e. g., SNPs with a small *MAF*, may lead to estimation problems as some genotypes may only be found in a few individuals in the given sample or not be represented at all. This means that, e. g., a *MAF* of 0.1 corresponds to $p_{aa} = \mathrm{MAF}^2 = 0.1^2 = 0.01$, in other words, 1 % of the population is

expected to be homozygote minor — in a sample of similar size to our study, with 354 patients, we would expect to find 3 to 4 such individuals. For example the MAF threshold of 0.05 used in our procedure corresponds to $p_{aa} = \text{MAF}^2 = 0.0025$, e. g., 0.25 % of the population is expected to be homozygote with the minor allele—in a sample of 400 individuals we would expect to find 1 such individuals. If the MAF were MAF 0.01, one would need a sample of approx. 10 000 individuals to find one. Since this requires a distinctly large sample size to find 1 individual, such variants need to be removed. They may lead to numerical complications, such as unreliable or misleading study results. By setting the limit $0.05 <$ MAF rare genotype variations are excluded from the study and erroneous results can be avoided.

The quality measures *info score* and *certainty* are generated by the imputation program to indicate confidence in the data. The info score is attributed to reference genotype information associated with the *MAF* and ranges between -1 and 1. The higher a SNP's info score, the more reliable the imputation has been performed. An info score value of -1 means that the imputation has been undefined, e. g., has not been calculated. The certainty indicates the average reliability of the best-guess genotypes. Setting the thresholds for info score at 0.8 and certainty at 0.9 data, from which no reliable imputation is achieved, can be removed.

To summarize, the selection of SNPs was based on this set of conditions:

$$\begin{array}{ll} \text{MAF} & > 0.05 \\ \text{info score} & > 0.8 \\ \text{certainty} & > 0.9 \end{array}$$

which are parameters that indicate a high confidence of the imputed data and give additional confidence for correct data.

Moreover, all duplicated SNPs were excluded and GTOOL was used for including patient information to the genotype data. GTOOL is a program which can be applied to easily modify and adjust genotype data Freeman and Marchini (2012). With genotype information of 392 individuals and 5 747 441 SNPs, we obtain dosage values corresponding to the expected minor and major allele counts for each SNP.

The dosage is calculated as the weighted sum of genotype probabilities:

$$\text{dosage} = \sum_{gt \in \text{genotypes}} w_{gt} \cdot p_{gt} \tag{6.4}$$

where probabilities (p) and weights (w) are shown below (a represents the minor, A the major allele).

genotype		aa	aA	AA
genotype probability		p_{aa}	p_{aA}	p_{AA}
weight		2	1	0
dosage	$\sum_{\text{gt}} =$	$2 \cdot p_{aa} +$	$1 \cdot_{aA} +$	$0 \cdot p_{AA}$

The weights count the number of minor alleles (a) in the genotypes. Therefore, the dosage value represents the probability to observe the minor allele for this SNP, p_a. It adds the three genotype probabilities into one single value which simplifies subsequent calculations.

For numbers of SNPs per chromosome after quality control see Table 6.3.

Chromosome 1	437 975	Chromosome 12	278 867
Chromosome 2	485 994	Chromosome 13	218 714
Chromosome 3	417 623	Chromosome 14	188 927
Chromosome 4	423 782	Chromosome 15	159 207
Chromosome 5	380 542	Chromosome 16	166 307
Chromosome 6	380 378	Chromosome 17	140 720
Chromosome 7	335 081	Chromosome 18	165 517
Chromosome 8	327 611	Chromosome 19	115 774
Chromosome 9	249 867	Chromosome 20	126 428
Chromosome 10	297 883	Chromosome 21	83 734
Chromosome 11	294 398	Chromosome 22	72 112
		Total:	5 747 441

Table 6.3: Number of SNPs per chromosome after quality control of the *TUM 1 dataset*.

This output eventually features correctly examined data to begin with calculations.

6.1.3 Genome-wide association study

A *genome-wide association study* (GWAS) is performed to analyze genetic variations of the genome to identify allele expressions which appear commonly together with a specific phenotype, e. g., a disease.

The GWAS can be calculated using PLINK – a program, that provides comprehensive analysis selections for genomic studies Purcell (2006); Purcell et al. (2007). Selected markers, e. g., SNPs, are chosen to be examined, most of them localized in non-protein-coding regions, such as introns or between genes. A possible association to a specific phenotype is evaluated, as shown in Fig. 6.2 on the next page. On the left side of the figure, there is no significant difference in the phenotype between the different genotype groups. The slope of the fitting line is rather flat. On the right side of the figure, a homozygote major allele (*AA*) shows a significantly different manifestation of the phenotype than the homozygote minor allele (*aa*). Since the heterozygote phenotype (*aA*) lies directly in the middle of the two homozygotes, this SNP's effect would be well described by a dosage model indicated by the steeper fitting line. For GWAS these associations for all known SNPs on the genome were computed, yielding a p-*value*, which indicates the significance of each finding.

The p-value represents the probability of finding a more extreme test-statistic than the observed one. It ranges between 1 indicating no and 0 indicating high significance. This means that the lower the p-value the less likely it is to find such data under the null-hypothesis and the more likely it is that the alternate hypothesis can be accepted. In standard clinical trials, the threshold is commonly set to 0.05 (5 % probability) for single tests, for genome-wide analysis to $5 \cdot 10^{-8}$ to account and correct for multiple testing Barsh et al. (2012). In case of deceeding the threshold, the result can be assessed as significant and the null-hypothesis rejected. For example, when having normally distributed data (with mean $= 0$ and standard deviation $\sigma = 1$,

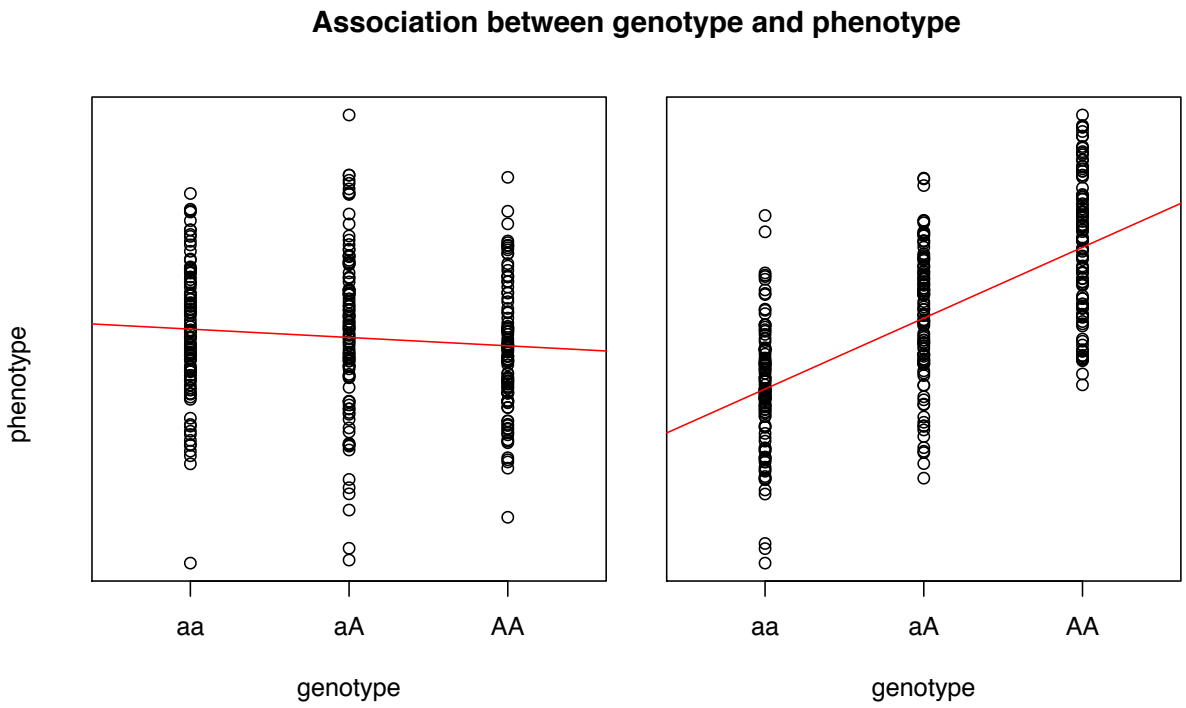

Figure 6.2: The figure on the left shows no, whereas the figure on the right shows a significant association between genotype and phenotype.

for which the 95 % confidence interval lies between −1.96 and +1.96, as shown in Fig. 6.3 on the following page) the probability of having a significant result above the 95 % confidence interval, within an one-sided test, is indicated by the red area in Fig. 6.3 on the next page which equals 2.5 % of the total area under the bell curve. If we consider all values outside of the confidence interval including also the area to the left of −1.96 (again 2.5 % area under the curve, AUC), corresponding to a two-sided test, we find 5 % in the marked region of the tails.

Using this data, the GWAS was performed on the measured antibody titer against interferon-β as well as on the normalized antibody titer. Both calculations were interesting because the antibody titer values are distributed differently as shown in Fig. 6.1 on page 45. The covariates *sex* and *age at sampling* (AAS) as well as components C1 to C5 from the multi-dimensional scaling (MDS) analysis of the identity-by-state matrix, which outline genotype similarity of a population, were included. Figure 6.4 shows an example of a scatter plot of the MDS components C1—C5.

The GWAS on the normalized phenotype data yielded some low p-values, nevertheless, no SNP reached the genome-wide significance limit. Our results showed 390 SNPs with p-values between 10^{-5} and 10^{-6}, the highest association found for SNP rs8051893 on chromosome 16 with a p-value of $3.515 \cdot 10^{-7}$. This SNP is localized in intron 1 of the WFDC1 gene, which is considered a tumor suppressor gene. All information of a genes localization and function are retrieved from the *Database of Single Nucleotide Polymorphisms*, dbSNP within this thesis Bethesda (2005); Sherry et al. (2001).

The resulting top SNP rs697296 of the secondly performed GWAS, our phenotype being the

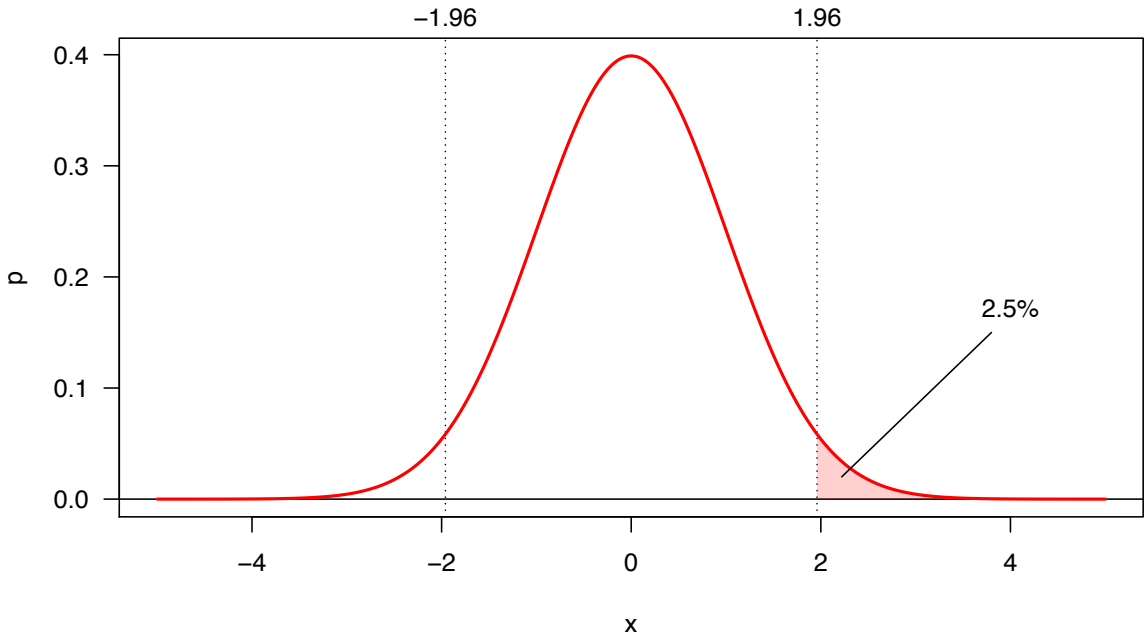

Figure 6.3: Interpretation of the p-value. The confidence intervall indicates the area between the two vertical lines marked at −1,96 and 1,96. The red area displayes the area under (the) curve (AUC) of 2.5 %, showing the significant domain.

SNP	allele 1	allele 2	frequency allele 1	info score	β	SE	p-value
rs8051893	C	T	0.5925	0.8885	0.373	0.07	$3.515 \cdot 10^{-7}$

Table 6.4: Top SNP from GWAS with normalized antibody titer of the *TUM 1 dataset*.

measured antibody titer ('AB') and also including covariates sex, age at sampling ('AAS') and C1 to C5, showed a p-value of 4.141 x10^{-7}.

SNP	allele 1	allele 2	frequency allele 1	info score	β	SE	p-value
rs697296	C	T	0.4169	0.9230	15.6294	3.02	$4.141 \cdot 10^{-7}$

Table 6.5: Top SNP from GWAS with measured antibody titer ('AB') of the *TUM 1 dataset*.

Although not directly localized on a gene, *dbSNP* reports its position close to the PRICKLE gene on chromosome 3. Its p-value is not significant, but including this particular SNP to the 386 already persisting SNPs in the PRICKLE gene, the SVM prediction model could yield higher correlation between measured and predicted antibody titer than without SNP rs697296. The r-value increased from 0.471 to 0.473 caused by the affect of only one additional SNP. For further details see chapter 5.

A *Manhattan Plot* provides an overview of the genome-wide p-values obtained from the GWAS. The genomic position of each SNP over all chromosomes is displayed along the x-axis and the negative logarithm of the p-value is on the y-axis. Each point represents the calculated p-value at the localization of one SNP. In Fig. 6.5 on the next page the top SNP on chromosome 3 with the lowest p-value is clearly recognizable.

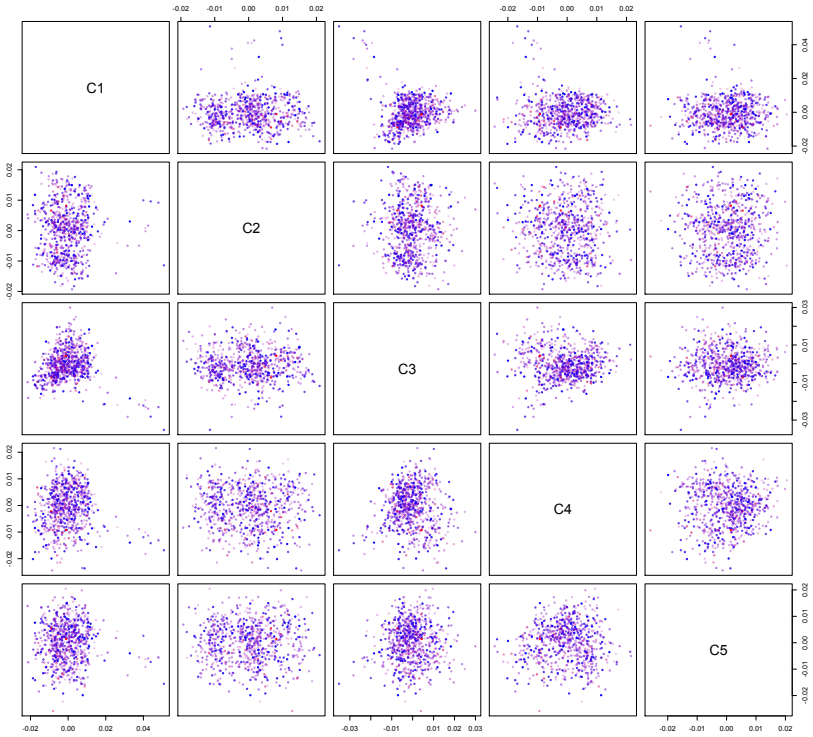

Figure 6.4: Scatter plot of the MDS analysis showing the components C1—C5.

Another way of interpreting the distribution of p-values in our study is by using the *Quantile-Quantile Plot* (QQ-Plot). The distribution of two variables are displayed to be compared. If all dots are disposed along the diagonal, their distribution would be equal. In this case the testing values are arranged according to the negative logarithm of the p-values of our data, a comparison of expected and observed p-values, as shown in Fig. 6.6 on page 53. Slight deviations can be accepted but the approximately comparable outcome of observed and expected p-values is an important requirement for continuing the study.

Carefully preparing our data and the lack of finding any significant SNPs correlating with our phenotype led us to search for advanced techniques to improve prediction beyond the single SNP results and also include possible interactions. One technique that allows this kind of analysis is machine learning with support vector machines. We continued our project with the intention of creating a model with SVM calculating SNP-interactions, see chapter 5.

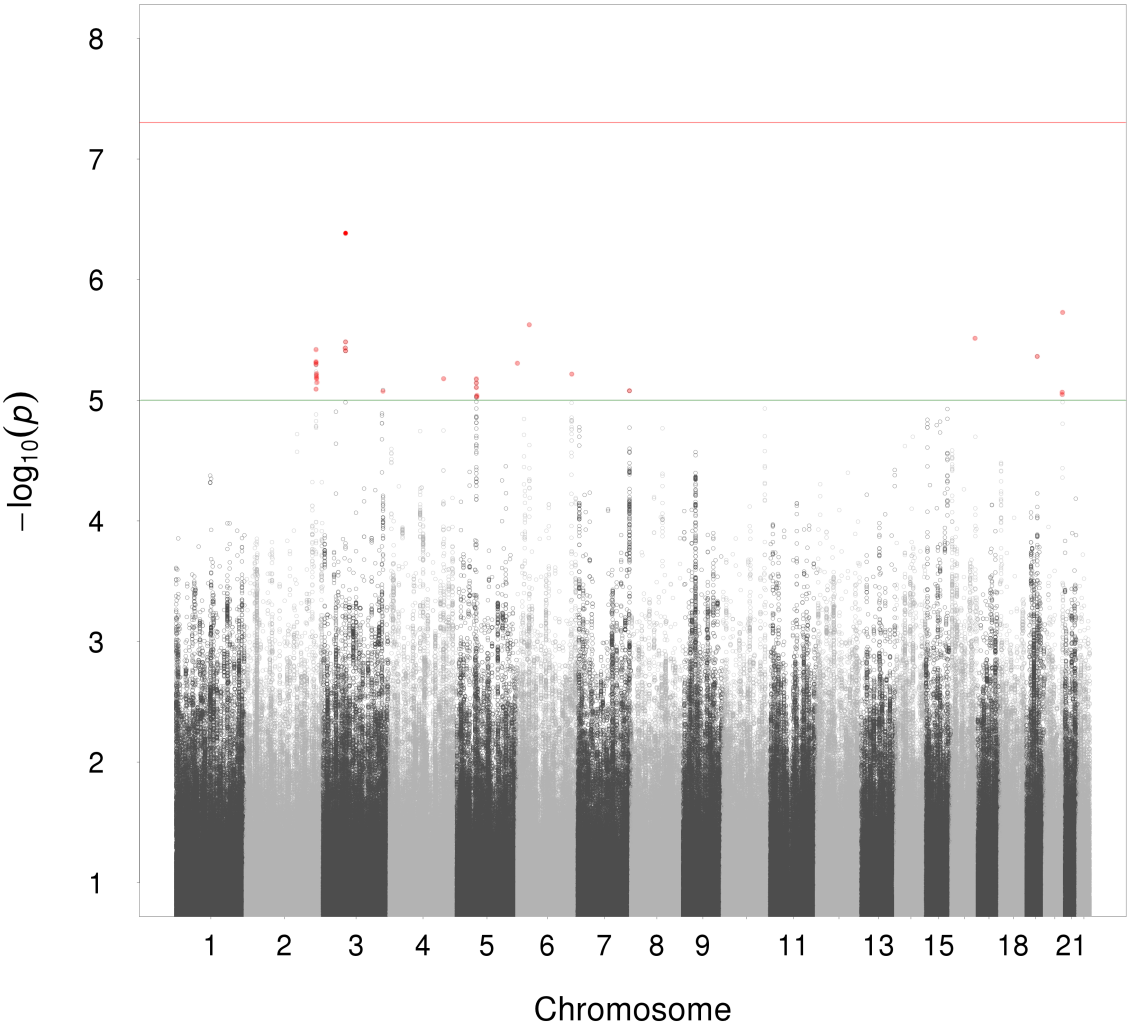

Figure 6.5: Manhattan plot of the *TUM 1 dataset* with the obtained p-values from the GWAS, the green line indicates suggestive, the red line genome-wide significance.

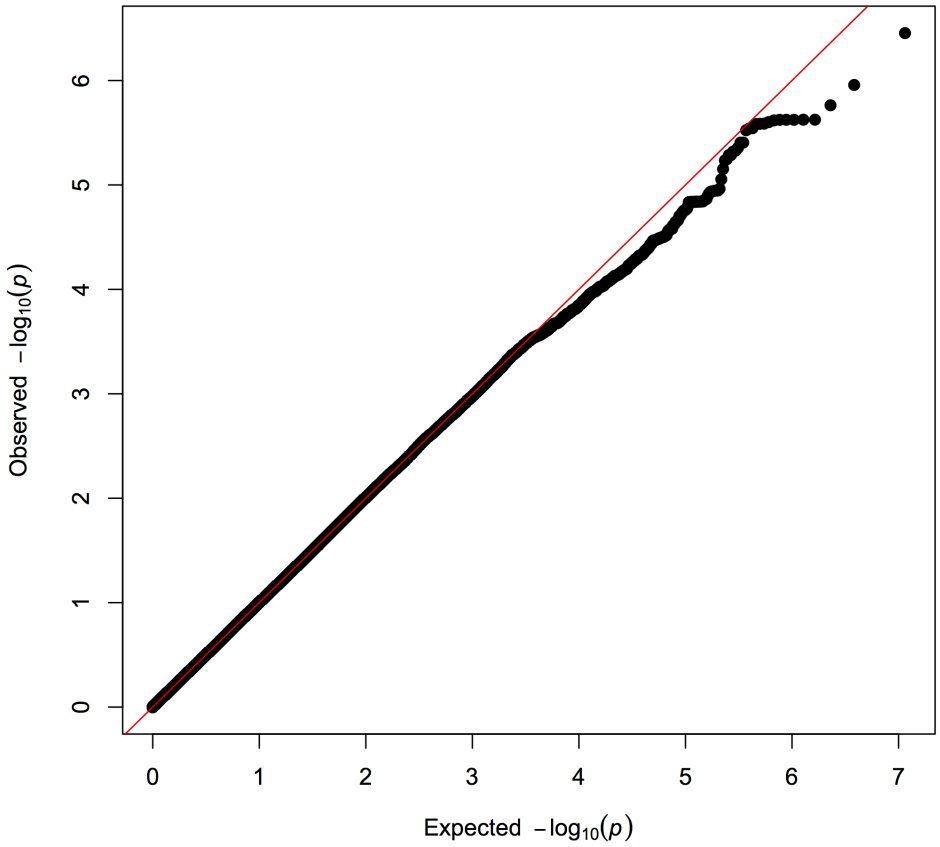

Figure 6.6: QQ plot of the *TUM 1 dataset*.

6.2 TUM 2 Dataset

The *TUM 2 dataset* is a larger sample of multiple sclerosis patients treated with interferon-β. We received the dataset from the Department of Neurology at the Rechts der Isar Hospital for this project. It contains genotype information of 1000 individuals — 272 males and 728 females, of which 762 have known medication and antibody status. The majority of 550 patients are antibody negative, whereas 113 patients develop neutralizing and 98 binding antibodies, as listed in Table 6.6. Note that in this study, a total of 41 percent of patients treated with Betaferon developed binding or neutralizing antibodies, nearly twice as many compared to other medications.

The *normalized* antibody titer is referred to a relative antibody titer in relation to a reference serum. The calculation was performed as follows:

$$\frac{AB - AB_{\text{ref}}}{AB_{\text{ref}}} \cdot 100\,[\%] \tag{6.5}$$

Medication	Patients	AB negative		BABs		NABs		unchar	
Betaferon	264	155	58.7 %	59	22.3 %	49	18.6 %	1	0.4 %
Avonex	188	170	90.4 %	8	4.3 %	10	5.3 %		
Rebif 22	79	58	73.4 %	10	12.7 %	11	13.9 %		
Rebif 44	211	153	72.5 %	18	8.5 %	40	19.0 %		
unknown	20	14	70.0 %	3	15.0 %	3	15.0 %		
Total	762	550		98		113		1	

Table 6.6: Antibody status by interferon-β medication of the *TUM 2 dataset*. For 20 patients the medication status was unknown: 14 cases showing no and 3 each developing binding or neutralizing antibodies. One patient treated with Betaferon is listed with uncharacterized positive antibody status. We did not exclude those cases from the study since we do not specify in subgroups of treatment.

Table 6.7 on the facing page shows the SNP count per chromosome after quality control, a total of 5 974 434 SNPs. The QC was performed using the same conditions as for the dataset *TUM 1 dataset*, for details see section 6.1.2 on page 45.

With this larger dataset, we again performed a GWAS analysis with the same conditions and covariates as described in section 6.1.3 on page 48 (with antibody titer as phenotype, and covariates being C1–C5 of MDS as well as sex and age of the samples). Evaluating the GWAS results we found two SNPs rs57658648 and rs57962245 as shown in Table 6.8 on the next page — both localized in close proximity of the HLA-DRB6 gene on chromosome 6 — with remarkable p-values of $8.508 \cdot 10^{-8}$ and $8.531 \cdot 10^{-8}$, respectively, just barely missing the genome-wide significance level. Since these SNPs were not present in the *TUM 1 dataset*, the HLA-DRB6 gene did not attract attention so far. Bearing in mind that many genes within the HLA region on chromosome 6 have influence on antibody production, this result of two almost significant SNPs may show importance for predicting treatment efficiency.

Chromosome 1	456 970	Chromosome 12	289 949
Chromosome 2	502 568	Chromosome 13	227 261
Chromosome 3	431 315	Chromosome 14	199 051
Chromosome 4	439 574	Chromosome 15	168 726
Chromosome 5	391 882	Chromosome 16	159 936
Chromosome 6	392 974	Chromosome 17	151 306
Chromosome 7	350 966	Chromosome 18	172 031
Chromosome 8	335 010	Chromosome 19	132 067
Chromosome 9	259 153	Chromosome 20	133 235
Chromosome 10	311 076	Chromosome 21	85 997
Chromosome 11	305 855	Chromosome 22	77 532
		Total:	5 974 434

Table 6.7: Number of SNPs per chromosome after quality control of the *TUM 2 dataset*.

SNP	allele 1	allele 2	frequency allele 1	info score	β	SE	p-value
rs57658648	G	C	0.9436	1.0208	-19.169	3.5	$8.508 \cdot 10^{-8}$
rs57962245	G	A	0.9436	1.0208	-19.166	3.5	$8.531 \cdot 10^{-8}$

Table 6.8: Top two SNPs from GWAS with normalized antibody titer of the *TUM 2 dataset*.

Moreover, in the GWAS of the *TUM 2 dataset* we found ten SNPs with p-values smaller than 10^{-6} as listed in Table 6.9.

SNP	chromosome	position	p-value
rs114125694	2	199 553 600	$9.494 \cdot 10^{-7}$
rs34349601	2	199 553 892	$8.819 \cdot 10^{-7}$
rs7578102	2	199 554 281	$8.436 \cdot 10^{-7}$
rs17826222	2	199 557 725	$3.963 \cdot 10^{-7}$
rs13410413	2	199 557 971	$3.973 \cdot 10^{-7}$
rs17826270	2	199 558 179	$4.026 \cdot 10^{-7}$
rs17826492	2	199 561 760	$6.816 \cdot 10^{-7}$
rs57658648	6	32 520 731	$8.508 \cdot 10^{-8}$
rs57962245	6	32 520 863	$8.531 \cdot 10^{-8}$
rs111875628	6	32 583 813	$6.536 \cdot 10^{-7}$
rs113414682	6	32 594 441	$4.317 \cdot 10^{-7}$
rs2537580	7	17 793 601	$3.199 \cdot 10^{-7}$

Table 6.9: Listing of SNPs with p-values less than 10^{-6} from the GWAS analysis on the *TUM 2 dataset*.

In comparison, the GWAS performed on the *TUM 1 dataset* of 354 individuals and 5 747 441 SNPs resulted 390 SNPs with p-values $< 10^{-5}$ the top SNP rs8051893 localized on the WFDC1 gene on chromosome 16 with a p-value of $3.515 \cdot 10^{-7}$ being the only resulting SNP with a p-value $< 10^{-6}$. These distinctly smaller p-values within the *TUM 2 dataset* can be explained by the larger sample count.

6.3 The combined sample (TUM 3 Dataset)

To evaluate the SVM model we decided to combine the *TUM 1* and *TUM 2 dataset* to achieve an even higher sample size for our final calculations. This combination of the samples was performed by including genotype samples of all (392 individuals) from the originally used *TUM 1 dataset* and only the non-overlapping part (804 individuals) of the *TUM 2 dataset*. Since the *TUM 2 dataset* consisted of not only new data, as some indviduals overlapped in both studies, we had to make sure to only include new and non-related individuals from the *TUM 2 dataset* to the merged dataset. This means that from the inital 1000 individuals from the *TUM 2 dataset*, a total amount of 194 patients were excluded because of duplication within the *TUM 1 dataset*. Moreover, 2 relatives were eliminated from the combined dataset as well, due to the strong resemblance of their genotype features. After these exclusions, 804 patients from the *TUM 2 dataset* remained in the combined study. Equally, not all of the included samples provided phenotype information. This resulted in 354 patients from the *TUM 1 dataset* as well as 728 patients of the *TUM 2 dataset* having phenotype information available.

Consequently, when adding up the results, 1196 individuals with genotype data and 1082 patients with phenotype information could be added to the combined dataset. Nevertheless, in process of the data merge some individuals again had to be dropped when no completeness of the covarites was given. 304 patients had to be removed due to unspecified sex, as also implemented within the GWAS-analysis. In conclusion, after merging the data, the dataset comprises 892 multiple sclerosis patients treated with interferon-β.

Again using this data, a QC was performed using the same conditions as for the *TUM 1 dataset* as well as for the *TUM 2 dataset*, for details see 6.1.2.

This resulted in a total number of SNPs slightly surpassing the 6 million count, which could be included in further calculations as listed per chromosome in table Table 6.10. Though the SNP rs4961252 on chromosome 8 does barely not pass quality control conditions, we decided to include this particular SNP for further calculations due to possible correlation with antibody production against interferon-β found in previous studies Weber et al. (2012).

Chromosome 1	463 598	Chromosome 12	294 316
Chromosome 2	510 168	Chromosome 13	230 904
Chromosome 3	437 159	Chromosome 14	202 238
Chromosome 4	446 311	Chromosome 15	171 205
Chromosome 5	397 827	Chromosome 16	179 402
Chromosome 6	399 560	Chromosome 17	153 601
Chromosome 7	356 193	Chromosome 18	174 545
Chromosome 8	340 477	Chromosome 19	133 588
Chromosome 9	263 006	Chromosome 20	135 260
Chromosome 10	315 773	Chromosome 21	87 349
Chromosome 11	309 774	Chromosome 22	78 523
		Total:	6 080 777

Table 6.10: Number of SNPs per chromosome after quality control of the combined dataset.

Using the GWAS-analysis strategy again with the phenotype indicating the antibody titer and covariates being C1–C5 of MDS as well as sex and age of the samples, we additionally included the study as a covariate to reflect a SNPs origin from the *TUM 1 dataset* versus the *TUM 2 dataset*. The study was initilized including 1196 individuals, whereas the analysis was performed with 891 patients, including all those presenting a non-missing phenotype. This means 304 individuals could not be included due to unspecified sex and one individual was excluded because of a missing alternate phenotype. As a result, a total of 6 SNPs on chromosome 6 yielded p-values $< 10^{-7}$ as listed in Table 6.11, in fact all localized in the proximity of the HLA-DRB1 gene on chromosome 6. This may indicate a correlation of antibody production against interferon-β and genes within the HLA-region.

SNP	allele 1	allele 2	frequency allele 1	info score	beta	SE	p-value
rs34958241	A	G	0.8492	0.7531	-13.69	2.43	$2.428 \cdot 10^{-8}$
rs34784936	G	T	0.8461	0.7448	-13.65	2.43	$2.457 \cdot 10^{-8}$
rs34855541	A	G	0.8502	0.7656	-13.42	2.42	$3.975 \cdot 10^{-8}$
rs35380574	C	T	0.8452	0.7793	-12.81	2.37	$8.382 \cdot 10^{-8}$
rs35395738	T	C	0.8523	0.8098	-12.88	2.38	$7.857 \cdot 10^{-8}$
rs35472547	G	T	0.8536	0.8175	-13.02	2.37	$5.383 \cdot 10^{-8}$

Table 6.11: Top SNPs from GWAS with normalized antibody titer of the combined dataset.

Moreover, 13 other SNPs localized on chromosomes 4, 6, 7, 13, and 15 yielded remarkable p-values of $< 10^{-6}$ and 92 SNPs yielded p-values of $< 10^{-5}$ listed in table Fig. 6.7 on the next page especially on chromosomes 6 and 13.

The resulting p-value of the promising SNP rs4961252 on chromosome 8, which has been associated with multiple sclerosis only yielded a value of $4.095 \cdot 10^{-5}$ in our study as shown in Table 6.12 Weber et al. (2012).

SNP	alleles	frequency allele 1	info score	β	SE	p-value
rs4961252	A G	0.6133	0.9252	-6.68	1.62	$4.095 \cdot 10^{-5}$

Table 6.12: GWAS result for the SNP rs4961252 within the combined dataset.

The Manhattan Plot, shown in Fig. 6.10 on page 61 nicely demonstrates the peak of low p-values on chromosome 6. The QQ plot, shown in Fig. 6.11 on page 62 shows an acceptable similarity of observed and expected p-value distributions. As a final step before employing the combined dataset for our SVM calculations we decided to compute the residual values of the antibody titer to be used as the phenotype. Residuals represent the variability of the data. In other words, they determine the deviation of the actual measured value from the estimated model.

SNP	*A1*	*A2*	*FRQ*	*INFO*	*BETA*	*SE*	*P-VALUE*
Chromosome 1							
rs12119103	G	A	0.8936	0.9338	-11.6556	2.5488	5.498e-06
Chromosome 2							
rs75821867	A	G	0.3770	1.9078	-26.1192	5.8328	8.524e-06
rs2565686	C	T	0.1981	1.3747	10.8874	2.4399	9.165e-06
rs28378381	A	G	0.5793	0.9745	-6.9411	1.5563	9.261e-06
Chromosome 3							
Chromosome 4							
rs7682820	G	C	0.4456	0.8954	-7.7023	1.6077	1.949e-06
rs9994029	G	A	0.4452	0.9040	-7.6098	1.6004	2.318e-06
rs10017348	C	T	0.4440	0.9073	-7.6746	1.5975	1.827e-06
rs9996749	G	T	0.4390	0.9074	-7.4367	1.6003	3.88e-06
rs4359979	C	G	0.4701	0.8873	-8.2148	1.6054	3.812e-07
rs4689374	G	T	0.4691	0.8904	-8.1259	1.6030	4.873e-07
rs4689375	T	C	0.4698	0.8891	-8.1497	1.6040	4.586e-07
rs7686248	T	A	0.4655	0.8860	-7.8897	1.6087	1.115e-06
rs13124547	G	A	0.4705	0.8963	-8.1965	1.5970	3.523e-07
rs13131705	C	A	0.4705	0.8976	-8.1892	1.5959	3.539e-07
rs12233714	G	A	0.4724	0.9048	-8.1762	1.5892	3.302e-07
rs76940812	G	T	0.3772	1.9144	-27.7519	5.6979	1.319e-06
rs72669292	A	G	0.3772	1.9146	-27.7591	5.6980	1.311e-06
Chromosome 5							
rs28405264	T	C	0.6480	0.9174	7.8132	1.6585	2.863e-06
Chromosome 6							
rs2395175	A	G	0.1139	0.9826	12.3395	2.4064	3.604e-07
rs34958241	A	G	0.8492	0.7531	-13.6908	2.4319	2.428e-08
rs34784936	G	T	0.8461	0.7448	-13.6487	2.4254	2.457e-08
rs34855541	A	G	0.8502	0.7656	-13.4154	2.4212	3.975e-08
rs35380574	C	T	0.8452	0.7793	-12.8089	2.3702	8.382e-08
rs35395738	T	C	0.8523	0.8098	-12.8804	2.3781	7.857e-08
rs35472547	G	T	0.8536	0.8175	-13.0151	2.3725	5.383e-08
rs34291045	A	T	0.8673	0.8561	-11.6732	2.4200	1.66e-06
rs34924558	C	T	0.8777	0.8505	-11.2109	2.5123	9.15e-06
rs34415150	A	G	0.8775	0.8514	-11.3442	2.5086	6.957e-06
rs34212923	T	C	0.8775	0.8517	-11.3511	2.5084	6.86e-06
rs34928543	G	C	0.8761	0.8515	-11.5540	2.4940	4.152e-06
rs34752364	G	A	0.8776	0.8659	-11.2656	2.4875	6.743e-06
rs36083025	A	T	0.8774	0.8645	-11.2586	2.4872	6.817e-06
rs2760976	C	T	0.9046	0.8224	-13.4065	2.8854	3.894e-06
rs35074855	C	G	0.9002	0.8310	-14.0098	2.7936	6.416e-07
rs35525122	C	A	0.8720	0.8544	-11.4184	2.4576	3.897e-06
rs17804379	C	A	0.9019	0.8139	-12.7638	2.8591	9.076e-06
rs35653258	C	A	0.8938	0.8372	-12.8842	2.7082	2.29e-06

Figure 6.7: Top GWAS results of merged datasets with p-values < $10 \cdot 10^{-6}$, 1.

SNP	*A1*	*A2*	*FRQ*	*INFO*	*BETA*	*SE*	*P-VALUE*
rs2647059	G	C	0.8953	0.8718	-12.6316	2.6628	2.446e-06
rs34039593	T	G	0.8677	0.8734	-10.6642	2.3965	9.691e-06
rs2647062	A	C	0.8733	0.8853	-12.3252	2.4140	4.04e-07
rs558721	C	T	0.8685	0.8744	-11.3094	2.3993	2.826e-06
rs679242	G	T	0.8583	0.8370	-11.1247	2.3770	3.316e-06
rs2647066	C	T	0.8652	0.8593	-11.1075	2.3947	4.044e-06
rs601945	A	G	0.8739	0.8515	-11.4717	2.4776	4.202e-06
rs617578	G	A	0.8934	0.8574	-12.3978	2.6665	3.836e-06
rs7761182	G	T	0.8704	0.8558	-11.6180	2.4377	2.197e-06
rs112485576	C	A	0.8719	0.8340	-11.3427	2.4888	5.907e-06
rs113881693	T	A	0.8714	0.8178	-11.6881	2.5082	3.65e-06
rs116753595	A	C	0.8722	0.8144	-11.7688	2.5203	3.487e-06
rs111344329	C	G	0.8669	0.8305	-11.6981	2.4478	2.063e-06
rs112397540	G	C	0.8653	0.8159	-11.6880	2.4573	2.302e-06
rs192602999	A	G	0.8642	0.8164	-11.5845	2.4483	2.593e-06
rs112969691	A	T	0.8636	0.8189	-11.5705	2.4402	2.473e-06
Chromosome 7							
rs2537575	G	A	0.7094	0.9556	-8.2044	1.7121	1.938e-06
rs2537580	C	T	0.7180	0.9600	-8.7550	1.7210	4.441e-07
rs2723525	G	A	0.7327	0.9478	-8.3687	1.7610	2.347e-06
rs2537583	C	T	0.7473	0.9784	-8.4021	1.7642	2.235e-06
rs2537584	A	T	0.7611	0.9655	-8.3445	1.8098	4.607e-06
rs2080060	G	C	0.7730	0.9733	-8.3740	1.8366	5.853e-06
rs1830004	G	A	0.7646	0.9510	-8.4893	1.8323	4.145e-06
rs2537589	G	A	0.7698	0.9729	-8.1299	1.8284	9.848e-06
rs2537590	A	G	0.7442	0.9580	-8.2184	1.7782	4.371e-06
rs17138250	G	T	0.7767	0.9842	-8.6309	1.8344	2.948e-06
Chromosome 8							
rs72692187	G	A	0.3763	1.9143	-25.3533	5.4064	3.173e-06
Chromosome 9							
Chromosome 10							
Chromosome 11							
Chromosome 12							
Chromosome 13							
rs147607590	A	G	0.9400	0.9711	-15.8890	3.2317	1.05e-06
rs77330495	G	A	0.9404	0.9709	-16.0100	3.2419	9.422e-07
rs192668777	G	C	0.9397	0.9634	-15.8322	3.2379	1.2e-06
rs75687330	G	C	0.9405	0.9744	-15.7157	3.2403	1.458e-06
rs79933822	T	C	0.9405	0.9744	-15.7119	3.2403	1.467e-06
rs41283964	C	T	0.9405	0.9750	-15.6692	3.2402	1.564e-06
rs41283966	C	A	0.9405	0.9748	-15.6487	3.2415	1.629e-06
rs41283968	A	C	0.9405	0.9748	-15.6459	3.2416	1.637e-06

Figure 6.8: Top GWAS results of merged datasets with p-values < $10 \cdot 10^{-6}$, 2.

SNP	*A1*	*A2*	*FRQ*	*INFO*	*BETA*	*SE*	*P-VALUE*
rs41283970	C	T	0.9405	0.9748	-15.6395	3.2419	1.656e-06
rs74092453	C	G	0.9405	0.9749	-15.6344	3.2420	1.67e-06
rs78968771	C	T	0.9408	0.9730	-15.5819	3.2534	1.962e-06
rs61118704	T	A	0.9407	0.9753	-15.5234	3.2457	2.027e-06
rs74092457	C	T	0.9409	0.9770	-15.3525	3.2498	2.689e-06
rs55723643	A	T	0.9410	0.9756	-15.3419	3.2551	2.834e-06
rs74092460	A	G	0.9411	0.9750	-15.3176	3.2588	3.013e-06
rs79899610	T	A	0.9411	0.9755	-15.3123	3.2594	3.048e-06
rs78878064	C	A	0.9412	0.9760	-15.2796	3.2612	3.237e-06
rs74441951	A	G	0.9410	0.9815	-15.0725	3.2472	3.982e-06
rs116469627	T	A	0.9408	0.9898	-14.7419	3.2313	5.777e-06
rs79244104	C	T	0.9408	0.9899	-14.7396	3.2311	5.792e-06
rs186649667	A	G	0.9409	0.9940	-14.6338	3.2275	6.582e-06
rs77532219	A	G	0.9409	0.9945	-14.6154	3.2266	6.719e-06
rs75663319	C	T	0.9409	0.9943	-14.6258	3.2272	6.648e-06
rs56347846	A	G	0.9409	0.9953	-14.5988	3.2262	6.863e-06
rs55815844	T	C	0.9409	0.9954	-14.5975	3.2261	6.873e-06
rs77983781	A	G	0.9409	0.9955	-14.5940	3.2260	6.904e-06
rs77667178	G	A	0.9409	0.9956	-14.5923	3.2260	6.92e-06
rs79116569	G	T	0.9409	0.9956	-14.5921	3.2260	6.922e-06
rs7993163	A	G	0.9409	0.9956	-14.5950	3.2262	6.9e-06
rs990613	C	T	0.9409	0.9953	-14.6028	3.2275	6.884e-06
rs990614	A	T	0.9409	0.9949	-14.6105	3.2270	6.787e-06
rs79249645	G	A	0.9409	0.9947	-14.6145	3.2273	6.762e-06
rs75118882	T	C	0.9409	0.9947	-14.6151	3.2274	6.757e-06
rs79983361	G	A	0.9409	0.9946	-14.6167	3.2275	6.747e-06
rs9542045	A	C	0.0590	0.9951	14.5991	3.2304	7.052e-06
rs1118693	T	C	0.0589	0.9951	14.5967	3.2320	7.15e-06
rs74477296	C	G	0.9411	0.9952	-14.5958	3.2321	7.161e-06
Chromosome 14							
Chromosome 15							
rs113178069	A	G	0.9445	0.7778	-19.6141	3.7498	2.111e-07
rs213150	C	A	0.0196	0.9205	30.2380	5.8796	3.335e-07
Chromosome 16							
Chromosome 17							
rs9899744	C	T	0.7877	0.9146	-8.8723	1.9395	5.457e-06
Chromosome 18							
rs1047363	C	T	0.3760	1.9034	-26.9512	5.7409	3.097e-06
Chromosome 19							
Chromosome 20							
Chromosome 21							
Chromosome 22							

Figure 6.9: Top GWAS results of merged datasets with p-values $< 10 \cdot 10^{-6}$, 3.

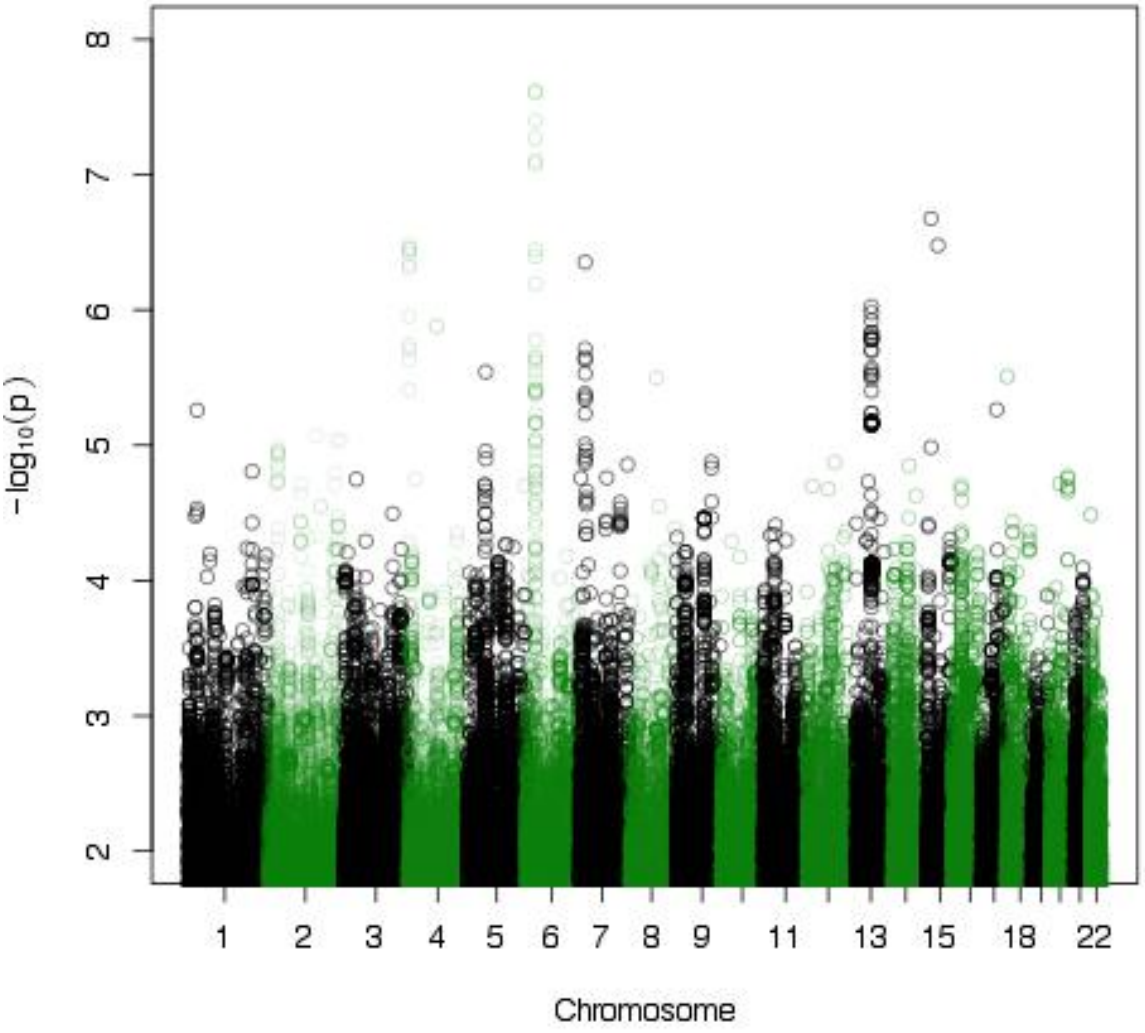

Figure 6.10: Manhattan plot of the GWAS results of the combined dataset.

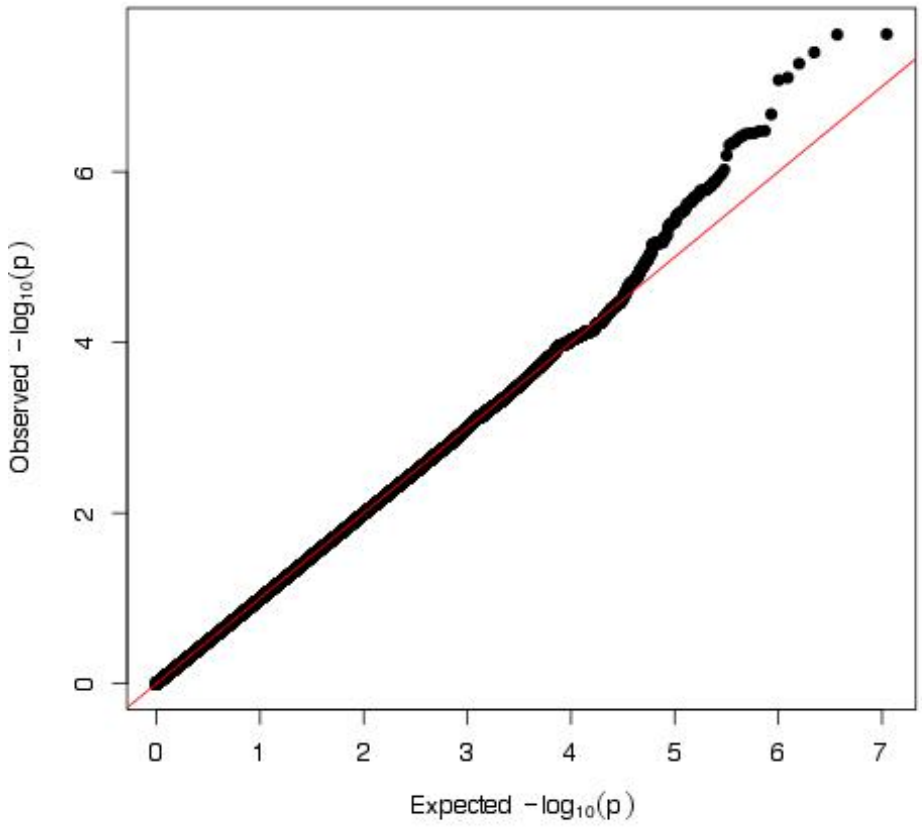

Figure 6.11: QQ plot of the GWAS results of the combined dataset.

7 Building prediction models from genotype data

7.1 Idea and setup

With today's improving research techniques, more and more genetic markers are being detected in association with developing various diseases, including the development of multiple sclerosis and other autoimmune inflammatory diseases.

Patients treated with interferon-β can produce neutralizing antibodies, which makes the medication uneffective. Immunologically relevant genes are found especially in the HLA regions on chromosome 6, where yet, some biomarkers have been associated with antibody production against interferon-β Barbosa et al. (2006); Buck et al. (2011); Buck and Hemmer (2014); Hoffmann et al. (2008); Link et al. (2014); Soelberg Sorensen (2008); Weber et al. (2012). The ability to create prediction models with SVM motivated us to use this approach to develop a method to forecast therapy response. Genotype information as well as antibody titer of patients treated with interferon-β provides us the data needed for SVM calculations. Finding a prediction to therapy response could be a major leap to avoid ineffective medication.

Before starting the calculations, it was important to explore the best possible preconditions for support vector machines to create a promising prediction model. These steps include

- finding the optimal SVM parameters — such as `cost` and `gamma`,
- finding the maximum SNP number support vector machines can calculate with using simulated data and prefilter correlated features
- dividing the genome-wide data into adjusted partitions — into single genes

and will be discussed in the following sections, which trace the development of the thesis and, therefore, present some intermediate results that cannot be generalized.

7.1.1 Finding optimal SVM-kernel parameters

When using SVM, it is recommended to first find the optimal parameters suitable for specific data to be analyzed. The function `auto.tune.svm()` in the **R**-package **e1071** implements a grid search over all used parameters. For a Gaussian kernel these are `cost` (C) and the kernel parameter `gamma` (γ), for details see section 5.6. *Auto-tunig* is able to detect the optimal parameters for the processed dataset based on the achieved performance.

We found only negligible differences between sampling methods leave-one-out cross validation (LOOCV) or n-fold cross validation (almost independent of n), so we usually employed 3-fold cross-validation for efficiency reasons.

Selecting the kernel The `svm` parameter `kernel` is, by default, set to *radial basis kernel*. The *radial basis kernel*, RBF, is what we called "Gaussian kernel" so far and generally used when smooth estimates are desired Smola et al. (1998). Another reason to go with this kernel is our goal to detect possible interactions between features (SNPs). Nevertheless, we performed calculations with different kernels listed in Table 5.1 on page 38 to find the best preconditions for our calculations. Overall, the *radial basis kernel* seemed to be the best option, see 7.2.2.3 on page 72.

This initial step to find a good choice of parameters helped us to later obtain reliable results in the study.

7.1.2 Estimation of confounding effects

There are another two types of effects that we needed to be investigated before actually working with the real data to know what kind of influence to expect on the resulting model. On one hand, SNPs that do not contain any relevant information related to the phenotype will likely reduce the predictive power of the model, as they unnecessarily increase the number of features. On the other hand, SVMs are known to poorly handle correlated input variables, such as SNPs with similar genotypes across the subjects.

7.1.2.1 Influence of non-informative features

To determine the maximum number of SNPs that support vector machines can reliably handle, we decided to create a genotype matrix with only one single real SNP from the data set and randomly simulated genotype data. These artificial SNPs were sampled from genotype distribution for a randomly chosen minor allele frequency (MAF) between 0.05 and 0.5, therefore containing no information which could positively influence the model.

We selected the SNP rs4961252 since previous studies on interferon-β therapy indicated a promising association with the antibody titer Weber et al. (2012). This study noted that even the best non-HLA single-SNP effect could account for about 2.6 % of the total phenotype variance, considered a rather small or weak effect. In other words, this means that 2.6 % of the anti-interferon-β antibody response can be attributed to this SNP.

For the prediction based on the SVM model with only this particular SNP plus *sex* and *age at sampling* as covariates, we find a correlation r^2-value of approximately 0.035 with the normalized AB titer. This is quite comparable to what is expected after the trial results mentioned above. We then started adding non-informative SNPs to increase the total number of SNPs from $n = 2, 5, 10, \ldots$ in approximately logarithmic steps to 5 000, see the labels on the x-axis in Fig. 7.1 on the next page which uses a logarithmic scale. For each SNP count, 7

simulated genotype matrices were generated. The resulting r^2-values, as displayed in Fig. 7.1, are where the violins indicate the distributions of the r^2-values per SNP count, while the respective mean r^2 is shown as circle with a smooth estimator line in red. As expected, the r^2-values steadily decrease with increasing numbers of random SNPs, which suggests that the random SNPs tend to be confounding factors for support vector machines.

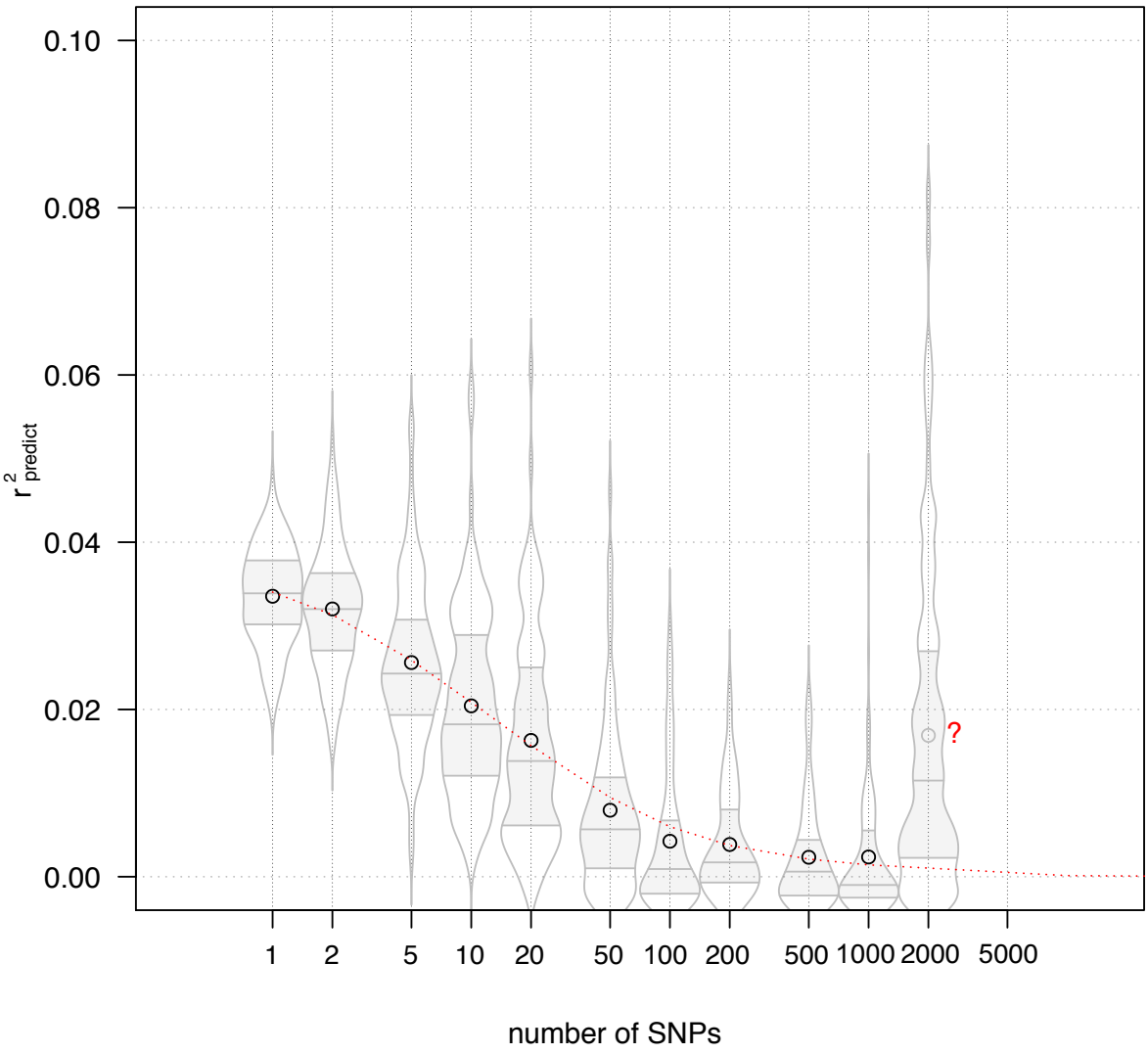

Figure 7.1: Exploring SVM predictability versus the number of SNPs included when all but one are non-informative, note the logarithmic scale on the x-axis. The dots indicate the mean of each 7 iterations with a smooth estimator line in red. The question mark indicates a possible effect of overfitting when adding 2000 or more non-informative SNPs to the calculations.

Adding a small number of SNPs the SVM could still identify our real SNP rs4961252 and yield an r^2-value comparable to the one found with no extra SNPs. At the limit of 500–1 000 random SNPs we observe r^2-values close to 0, which determined our decision to not include more than a few hundred SNPs per SVM model. Furthermore, by including over 2 000 random SNPs we can again observe an increase of the r^2-values. This may be caused by *overfitting*,

which means that too much information is given, so that in any situation a considerable but meaningless prediction result can be achieved. At $n = 5000$ we find r^2 close to 1, which is not included in the plot. However, the SNP data in this test were simulated as independent, which means without any correlation structure. In real genotyping data, we have to expect some degree of LD in between the SNPs, as described in 7.1.2.2, which will lead to a faster decrease of performance. In other words, for reliable results one should probably aim at SNP numbers well below those estimated from Fig. 7.1 on the preceding page when building a model from real data.

7.1.2.2 Correlated features

The effect of correlated genotypes in the genotype matrix can be estimated with a similar approach. With the function `preFilter()`, using the function `findCorrelation()` available in the **R** package **caret** Kuhn et al. (2016), the correlation threshold can be chosen and thus the correlation of SNPs evaluated.

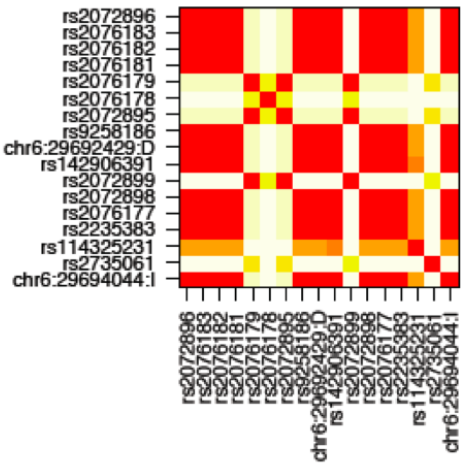

Figure 7.2: Pairwise correlation of 17 SNPs localized on the HLA-F gene. The red colored squares indicate haplotypes, the color attenuating with the degree of correlation.

Looking at all pairwise correlations (see Fig. 7.2) of SNP genotypes from, e. g., one gene, it is common to find *linkage disequilibrium (LD)* blocks which were introduced in 6.1.2. As a practical consequence, only one SNP within a LD block of highly correlated SNPs, also called haplotypes, needs to be kept in the study. Haplotypes denote a certain genotype combination of several SNP, which guide the formation of LD blocks.

Since we do not have access to the haplotype information for our genotyped patients, we take the correlation between the genotypes as a proxy and assume that highly correlated SNPs ($r > 0.9$) can be considered equivalent to an LD block and all but one SNP are excluded from the primary analysis. This way, we can in some cases drastically decrease the number of SNPs localized on one gene that need to be considered in the analysis and achieve significant time savings for the calculations. Setting the correlation cut-off to $r > 0.75$ the number of remaining SNPs can be reduced even further, which turns out to be necessary for larger genes with more than 1 000 SNPs.

7.1.3 Dividing data into adjusted partitions

SVMs can only generate reliable prediction models when a certain number of given variables is not exceeded, see 7.1.2 on page 64. Since we work on a data set with over 5 700 000 SNPs, it was necessary to split our data into more suitable partitions for the SVM calculations. One prediction model should likewise be created with an optimal number of SNPs of not more

than a few hundred. Therefore, we decided to create one prediction model for each gene. This means that all SNPs localized on one gene will supply genotype information for one calculation. Later on, all models can be summarized to yield a genome-wide model for therapy outcome prediction.

7.2 Building prediction models from genotype data

This section focuses on the implementation of the method on the real genotype dataset. Fig. 7.3 shows an overview of our project concept, setup, and idea, which will be introduced step by step in the following sections. For building the prediction model the *TUM 1 dataset* was used.

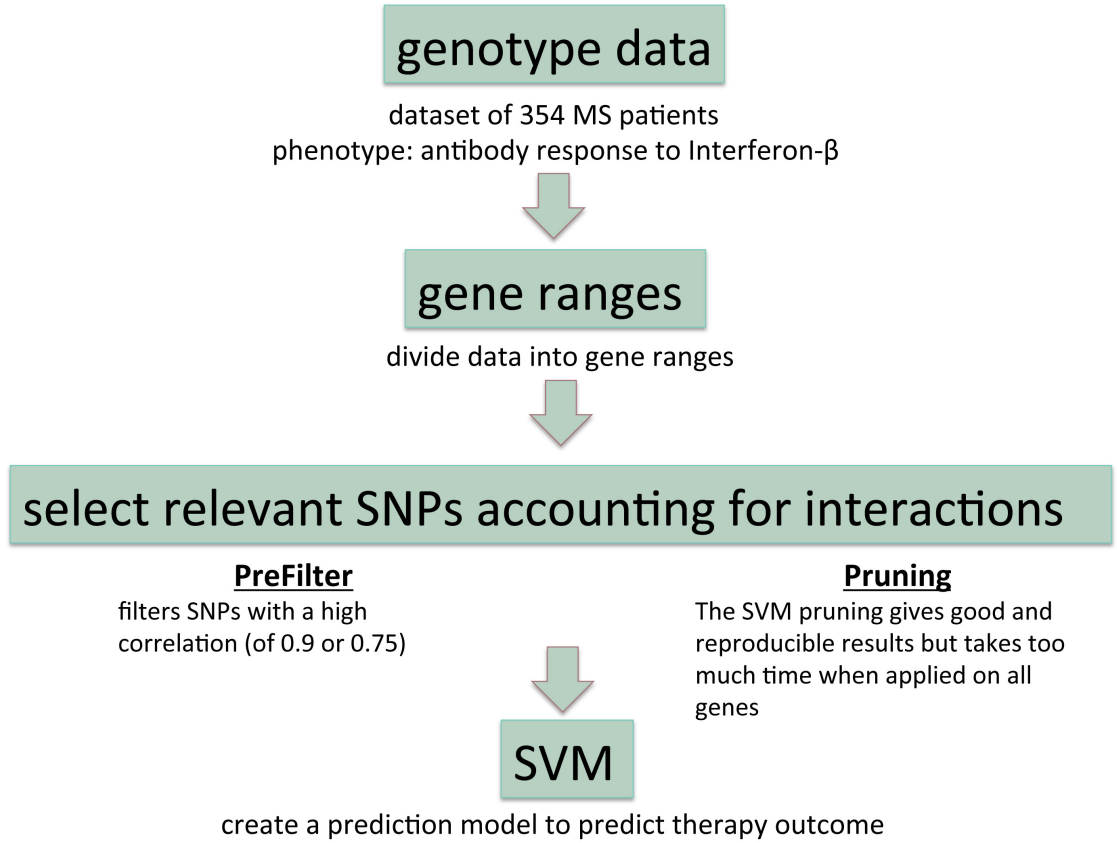

Figure 7.3: Project concept.

7.2.1 Finding gene ranges

Due to the high number of over five million SNPs in the study, it was necessary to divide the data into more suitable partitions for SVM calculations. As demonstrated earlier in our study, SVM predictability gets unreliable because of *overfitting* when including 2 000 or more variables as shown in Fig. 5.1 on page 32. This led to the idea to partition the data into gene ranges and creating one SVM model with all SNPs localized within one gene or gene range.

A *gene range* was defined at the start base position of a gene, including all overlapping genes along the DNA strand, concluding with the end base position of the last overlapping gene. In other words, all SNPs localized on genes having overlapping positions on the DNA strand will be added to a combined SVM prediction model.

The genomic reference data from the *UCSC annotation database* provides information of in total 44 419 genes with their respective start and end base pair position on the reference genome Kent et al. (2002). We used the annotation database for the 2009 assembly of the human genome (hg19). A newer version was released in March 2014. The annotations were generated by the University of California at Santa Cruz (UCSC) and collaborators worldwide, for details see www.genome.ucsc.edu. This allowed us to select all known SNPs within a gene's boundaries. We again obtained the relevant references from *dbSNP* Bethesda (2005); Sherry et al. (2001).

While this resource is available online, for performance reasons it is advantageous to have that information readily available. We stored the data in a local relational database, in *SQLite* for testing whereas *PostgreSQL* for the full dataset. These databases and their respective interface packages to **R** **RSQlite** Müller et al. (2017) and **RPostgreSQL** Conway et al. (2016) gave us permanent and fast access to genome and SNP information and let us select relevant genes by SNP ID, base position, expected allele frequency as well as info score and certainty of each SNP, see section 6.1.2 on page 45.

This approach of partitioning the data, resulting in 21 772 gene ranges, would reduce the calculations of prediction models more than half than needed for each of the 44 419 genes. However, we not only found gene ranges of only one or a few genes per gene range, we also found some gene ranges with a high number of overlapping genes, up to 77 genes per gene range, were composed. This would also drastically increases the number of SNPs per calculation. Large gene ranges would contain clearly more than 1 000 SNPs, which is not aspired for SVM as identified in section 7.1.2.1. This fact seemed contrary to the aim of reducing the number of features in SVM calculations to avoid overfitting. So we decided to continue the study preferably creating one prediction model per gene.

Through this approach of partitioning the dataset into gene-wise subsets, we were able to reduce the number of SNPs per calculation, which made first calculations with SVM possible.

Knowing specifically that the HLA regions are associated with antibody production, we initially focused on chromosome 6, where the HLA genes are localized. For each gene on chromosome 6, we performed SVM calculations (using a three-fold cross validation, parameter `cost` being set to 1, `gamma` to 0.001, see section 5.6 on page 37). The predictive power of the model is expressed by the correlation r^2 between measured and predicted AB titer. Genes with high r^2 are potential candidates for carrying SNPs that could serve as predictive biomarker. Fig. 7.4 on the next page summarizes performances for the 2 101 genes on chromosome 6. The dots in the plot indicate the prediction r^2 of all genes on chromosome 6 along with corresponding number of SNPs in those genes exceeding r^2 of 0.1. The peaks of the shaded area in the background represent hotspots of interesting genes. The peak, around 40 Mb, nicely corresponds to the fact that antibody production seems partially controlled by HLA genes.

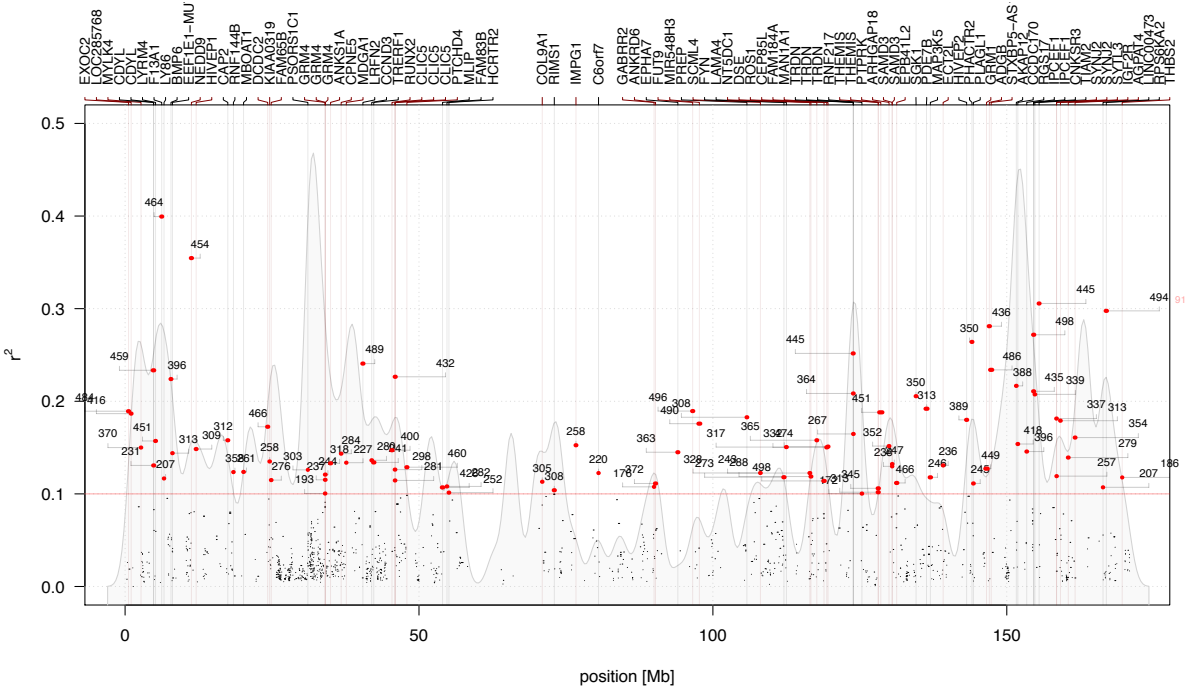

Figure 7.4: SVM performance of gene-wise calculations for chromosome 6. Calculations were performed using three-fold cross validation, parameter `cost` being set to 1, `gamma` to 0.001. The red dots indicate the prediction r^2 of the genes on chromosome 6 along with corresponding number of SNPs in those genes exceeding r^2 of 0.1, showing the names on the top. The peaks of the shaded area in the background represent hotspots of interesting genes. The peak around 40 Mb indicates the HLA region.

It is likely to find indicative SNPs also outside the proper gene boundaries, since they may be localized in promoter regions or otherwise have impact on the activation and regulation of a gene. Therefore, we decided to redo the SVM calculations with expanded gene boundaries of $\pm$200 kb.

7.2.2 Gene processing

PreFiltering is the first step to reduce the number of SNPs by eliminating highly correlated features, as introduced in section 7.1.2.2. Thereafter, SVM feature selection can detect relevant biomarkers by *pruning*. After these calcluation, we decided once more to optimize the SVM parameters *cost* and *gamma* as well as the chosen *kernel*, which will be presented in the following section.

7.2.2.1 Prefiltering

As introduced earlier in section 7.1.2.2, prefiltering is an effective method to reduce the number of features by excluding highly correlated SNPs from the study.

The largest gene on by number of SNPs on chromosome 6 is PARK2, for which a total of 4 235 SNPs is represented in this study. A mutation in PARK2 is associated with the development of Parkinson's Disease. When creating a prediction model with all SNPs, SVM (calculation of

a three-fold cross validation with `cost` set to 1 and `gamma` to 0.001) can yield a prediction power of 0.978. After prefiltering the SNPs on correlation of at least 0.9, which resulted in 574 SNPs, SVM can achieve a predictive value of 0.818. By prefiltering with a threshold of 0.75 correlation we were able to reduce the number of SNPs down to 257 and the r-value consequentely decreased to 0.631. This shows, as listed in table 7.1, that the r^2-value increases as the number of features adds, which indicates a good example of overfitting the model. Too many features can raise the prediction power.

preFilter	SNP count	r-value
—	4 235	0.987
0.9	574	0.818
0.75	257	0.631

Table 7.1: Prediction power in dependance of SNP count of the PARK2 gene.

Looking at two of the MHC II immune reaction associated genes in the HLA region, we find an increasing r-value as we prefilter the SNPs.

Within the SVM calculations (a three-fold cross validation with `cost` set to 1 and `gamma` to 0.001) of the HLA-B gene, a r-value of 0.241 can be achieved when adding all 82 SNPs to the dataset. After prefiltering with a threshold of 0.9 and 42 SNPs remaining an enhancement of prediction power up to 0.268 can be observed.

With a total of 45 SNPs localized on the HLA-C gene, the SVM calculation yields a prediction power of 0.177. After 0.75 prefiltering and having only 4 SNPs remaining, the r-value increases to 0.193. This may be an example of a gene, which contains relevant predictive information and yields higher prediction power when confounding features are eliminated.

	HLA-B		HLA-C	
preFilter	SNP count	r-value	SNP count	r-value
—	82	0.241	45	0.177
0.9	42	0.268	8	0.193
0.75	23	0.210	4	0.193

Table 7.2: Prediction power in dependance of SNP count of the HLA-B and HLA-C genes.

7.2.2.2 Feature selection: Pruning and growing approaches

SVM feature selection detects relevant biomarkers, such as SNPs or genes, which are selected for calculations and isolated from noisy data. There are two different straight-forward approaches to feature selection.

Pruning The process of starting with the whole feature set and eliminating one feature at a time is referred to as *backward elimination*. Pruning is performed within the usual SVM approach. The first SVM calculation includes the entire dataset and is used to create a prediction model. The features then are evaluated. The iterative feature selection process selects SNPs by

their importance with respect to the prediction outcome. The feature with the least contributing influence will be dropped. Another prediction model is created using the remaining (all minus one SNPs) and again the least contributing feature is dropped. After eliminating one feature at a time, this procedure is repeated until all features have been dropped. The last feature to be dropped indicates the feature with the most influence on the prediction outcome. This method of feature selection is also called *pruning*.

Algorithm 1 (Pruning)

1. *Start with all N features*
   2. *Build and evaluate a prediction model N times with always one feature removed, record performance*
   3. *exclude feature that was removed in the resulting model with best performance, this feature contributed least to a prediction with high performance in this round*
   4. *repeat from 2. until only one feature is left*

Growing The other approach works in the opposite way. The process begins with zero features, with each step one feature will be added to the model. This means the first SNP to be included for the study is the feature with the most contributing effect and prediction outcome is only based on its single SNP effect. Predictions models are then created by continuously adding one of the remaining SNPs. The SNP pair achieving the best performance is kept in the study. If the added SNPs contributed influence this would subsequently increase prediction power of the SVM model. The procedure will be repeated until all features are included. This method is also referred as *growing*.

Algorithm 2 (Growing)

1. *Start with zero of N features*
   2. *Build and evaluate a prediction model N times with always one feature, record performance*
   3. *Keep the feature that created the model with best performance, this feature contributed most to a prediction with high performance in this round*
   4. *Repeat from 2. until all features are included*

As an example for *pruning* we initially reviewed the HLA-DRB1 gene. Since the gene does not contain any SNPs after QC-filtering, we expanded the gene region by 10 kb, to find 101 SNPs within our newly set limits. The plot Fig. 7.5 on the following page traces the SNP elimination process versus achieved performance for the HLA-DRB1 gene. The local maximum at the right, indicating the optimal prediction, is achieved based on information of only 10 SNPs from this gene.

Using this test model, this gene — and many others independently — already yield improved prediction over single SNP approaches. These early results gave us a promising perspective that combining the genotypic information across genes and, eventually, the whole genome, should yield additional predictive power.

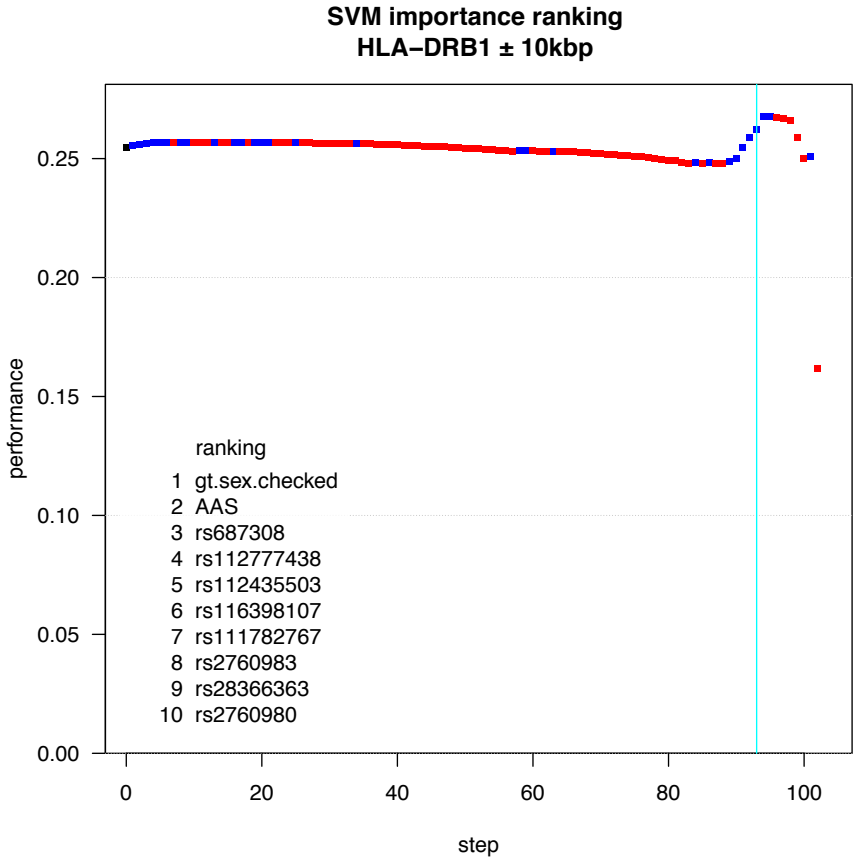

Figure 7.5: Pruning plot of the HLA-DRB1 gene. The SVM calculation was performed 101 times, for each step (*x-axis*) the performance (*y-axis*) is displayed. The maximum performance on the right side of the plot indicates the best performance including 10 SNPs, defined by the blue vertical line.

Practical considerations Apart from the encouraging and reliable performance of *pruning*, one should not forget one downside of these techniques: they have considerable computational requirements. Since a large number of SVM models are created within each iteration to compare and detect the most informative SNPs, a lot of calculation time and processors are needed.

7.2.2.3 Re-evaluating SVM parameters

As introduced in section 5.6, `gamma` (γ) is a *kernel* parameter in SVM calculations, see Table 5.1 on page 38. It determines how flexible a SVM model can respond to the data points in the fitting process.

Fig. 7.6 on the facing page reveals how far the data can spread depending on `gamma`. When `gamma` is set low, a variables influence reaches further. This means the smaller the `gamma`, the flatter the prediction curve. On the other hand, when `gamma` is set high, it will keep a feature's influence constrained, which allows the prediction to closely follow any feature in the data. Similarly, increasing `cost` (C), will force the prediction to more closely follow the data points. It is reasonable to believe that the extremes in either parameter are bad choices for building

meaningful models. If gamma is set too low, the prediction is forced to be flat and is unable to show fluctuations in the data. However, high gamma values lead to predictions that will perform perfectly on the given (training) data but are unlikely to perform equally well on new data. If there is no penalty associated with misclassification, which equals a small cost, a flat prediction results while a large cost leads to wild overshooting and inaccurately representing the training data. When aiming for a prediction model that on one hand provides a reasonable fit to the training data but can on the other hand be seen trustworthy when predicting the outcome for new data, a parameter combination of e. g., $C \approx 1$, $\gamma \approx 1$ seems appropriate.

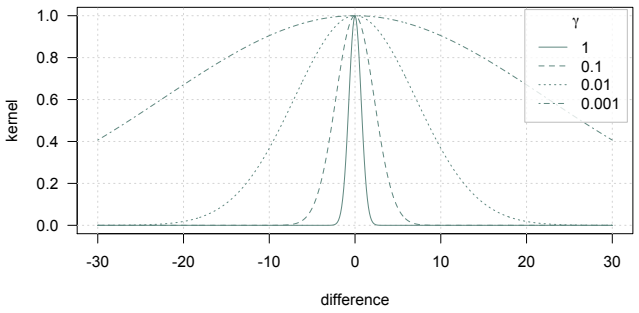

Figure 7.6: The SVM parameter gamma defines the reach of data. A high gamma keeps response constrained, small gamma allows further reach of model response.

We were concerned that we might have set the gamma to high, which could suppress single SNP effects. Some calculations showed that setting cost and gamma constant made it difficult to achieve high r-values.

Initially, gamma was determined by a cross-validation grid search for a medium size gene and held constant, irrespective of the model size. In retrospect, the chosen value $\gamma = 0.001$ turned out to keep the modeling process from adapting to the data and was able to achieve only poor performances, especially for small numbers of SNPs.

This made us re-evaluate the parameter and review calculations when gamma is calculated as a function of n, indicating the number of SNPs in the model:

$$\gamma \mapsto \gamma(n) = \frac{1}{n} \tag{7.1}$$

As presumed, these calculations are able to yield higher prediction power. The genes presented in section 7.2.2.1 achieved exemplary higher prediction power. Calculations with the PARK2 gene yield a higher r-value of 0.928, prefiltered at 0.9 with 574 SNPs, compared to an r-value of 0.818 when gamma was set to 0.001.

gamma	preFilter	SNP count	r-value
0.001	—	4 235	0.987
0.001	0.9	574	**0.818**
0.001	0.75	257	0.631
$1/n$	0.9	574	**0.928**

Table 7.3: Prediction power of the PARK2 gene with newly set gamma. The table shows an increase of the r-value when gamma is set to $1/n$.

Even more impressive is the notable increase of prediction power within the HLA-B gene. Compared to a constant gamma, calculations can now reach a more than two-fold increase of prediction power, as illustrated in Table 7.4 on the next page.

Fig. 7.7 shows the pruning plots for the HLA-B gene. The light blue curve indicates the pruning performance when gamma is set to 0.001. With a variable gamma, as defined in Eq. 7.1 on the previous page, the performance increases impressively as displayed on the y-axis of the pruning plot. This shows that a high gamma keeps the response localized and might be too susceptible to random fluctuations in the data. On the other hand, when gamma is too low, the response can no longer follow the data. Both effects lower the performance.

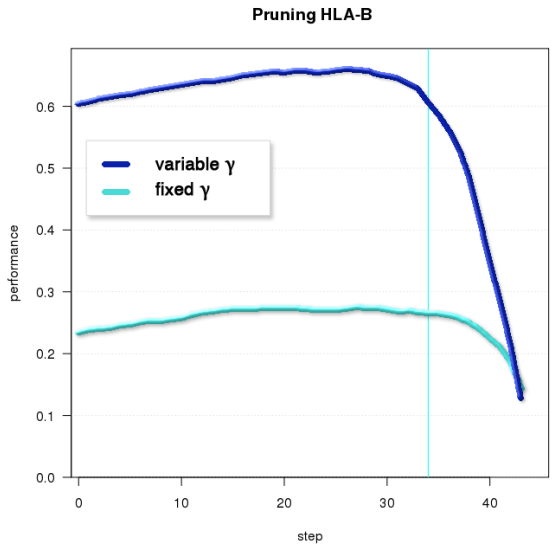

Figure 7.7: Pruning plot for the HLA-B gene in dependance of modified gamma. For each pruning step (*x-axis*) the performance (*y-axis*) is displayed. A clear gain in performance can be achieved when gamma is set to $1/n$.

		HLA-B		HLA-C	
gamma	preFilter	SNP count	r-value	SNP count	r-value
0.001	—	82	0.241	45	0.177
0.001	0.9	42	0.268	8	0.193
0.001	0.75	23	0.210	4	0.193
$1/n$	0.9	42	0.66	8	0.425

Table 7.4: Prediction power in dependance of gamma of the HLA-B and HLA-C genes.

Also the HLA-C gene calculations show a distinct improvement of the r-value from 0.193 to 0.425 within SVM calculations, see Table 7.4.

Secondly, we decided to also re-evaluate the kernel for further calculations. Figure 7.8 shows the performance of the pruning calculation of the HLA-DRB1 gene (gene boundaries $\pm$10 kb including 101 SNPs) using different kernels. At first glance, it seems like the higher the polynomial degree, the better the model performance gets. Approximately 20 features contribute influence to reach the maximum performance. Upon further inspection, we based our decision of kernel choice on the steepest ascent (from the right of the plot) indicating the

performance increase when adding features to the model. Therefore, the *radial basis kernel* seems to be the best choice.

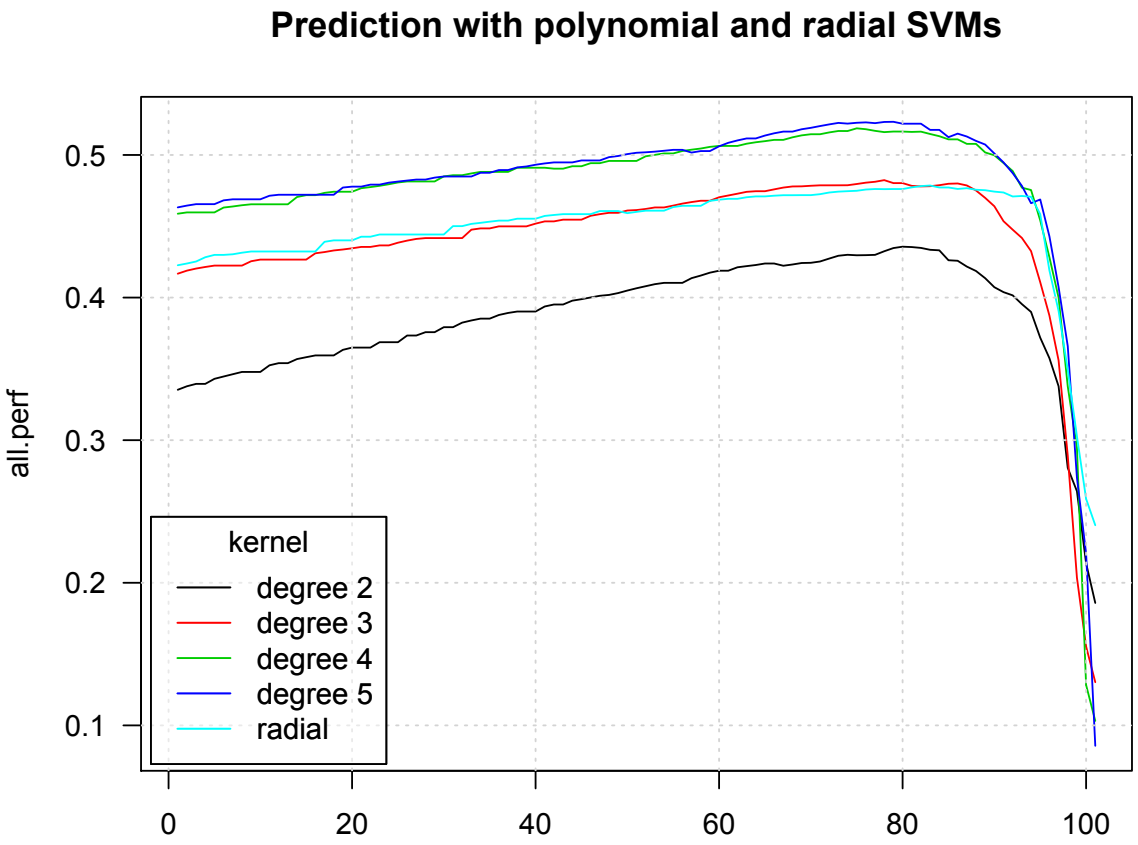

Figure 7.8: Comparison of pruning results of the HLA-DRB1 gene (gene boundaries ±10 kb including 101 SNPs) with different kernels set for SVM calculations. For each pruning step (*x-axis*) the performance (*y-axis*) is displayed. The pruning curves show calculations using different polynomial kernels degree 2 to 5 as well as the radial basis kernel. The best choice of kernel for calculations within this project is the radial basis kernel, as it shows the steepest ascent, on the righ side of the plot, when adding features to the model.

7.2.2.4 Permutation

Permutation is a well-known technique used to validate study results and review their significance. A permutation estimates the distribution of the null-hypothesis and is used to reproduce no significant effects.

Within a permutation of our project, we randomly assign phenotype information to different individuals to break up any real relationship between genotypes and phenotype. In other words, the genotype data are kept fixed while the antibody titer against interferon-β is randomly rematched to the individuals. Assuming there had been a meaningful association in the original

data, this would most likely have been destroyed in the new rearrangement of genotypes to phenotype. Calculated correlations for this situation should be weaker than for the real data. However, as discussed in section 6.1.3 in the context of calculation p-values within the GWAS, even for random data (under the null hypothesis) highly significant results may be found (false positives), as shown in Fig. 6.3 on page 50, although unlikely. In this sense, using a permutation produces data according to the null hypothesis. By repeating this step many times and evaluating the correlation each time, we get an approximation to the distribution of the expected correlation for random data. Comparing the correlation found for the real data with this distribution, the significance of the finding can be estimated. In some cases, a slightly better outcome can be accepted, such as in situations where data is, by chance, a better match than the real distribution.

For this reason, permutations need to be repeated many times to find a sufficiently fine approximation to the null distribution. Permutations should be repeated 100, 1 000 ... or even 10 000 times to avoid accepting an occasional outstanding result. This way, the permutation can represent the averaging distribution.

In order to evaluate the performance of the SVM pruning calculations, we performed a permutation of the phenotype values of the HLA-DRB1 gene $\pm$10 kb including 101 SNPs. The top plot in Fig. 7.9 on the facing page shows the permutation results. Once the phenotype was permutated an SVM model (cross-validation with newly set parameters) was created and evaluated. The blue line shows results of the real data, the x-axis indicating the number of included SNPs displayed against its r-values. For each number of SNPs (1 to 12) 100 permuations were performed with the resulting distribution of correlation values shown as violins. The numbers on the upper margin of the plot indicate how many permutation results achieved higher prediction than the real data e. g., the significance level in percent. It is noticeable that with only one SNP in the model, SVM yields the highest r-value with real data and no permutation can exceed this result. This indicates that the top SNP is a relatively powerful predictor within the HLA-DRB1 gene. However, when growing the model to include more SNPs the significance drops.

The plot on the left side in figure 7.9 shows permutation results (1 000 iterations) of the HLA-C gene as light grey violins in the background. This figure shows that permutations of different genes reveal similar results, indicating that SVM is able to achieve a certain amount of prediction power by creating the prediction model on the given training data, regardless whether influencing SNPs are involved. On the right side, permutation results of 10 genome-wide randomly selected SNPs, again with 1 000 iterations, were added. As expected, the high variance of the SNPs are able to achieve even higher r-values, whereas within the single SNP effects the randomly selected dataset can not outperform the real data and shows comparable results to the HLA-genes permutation.

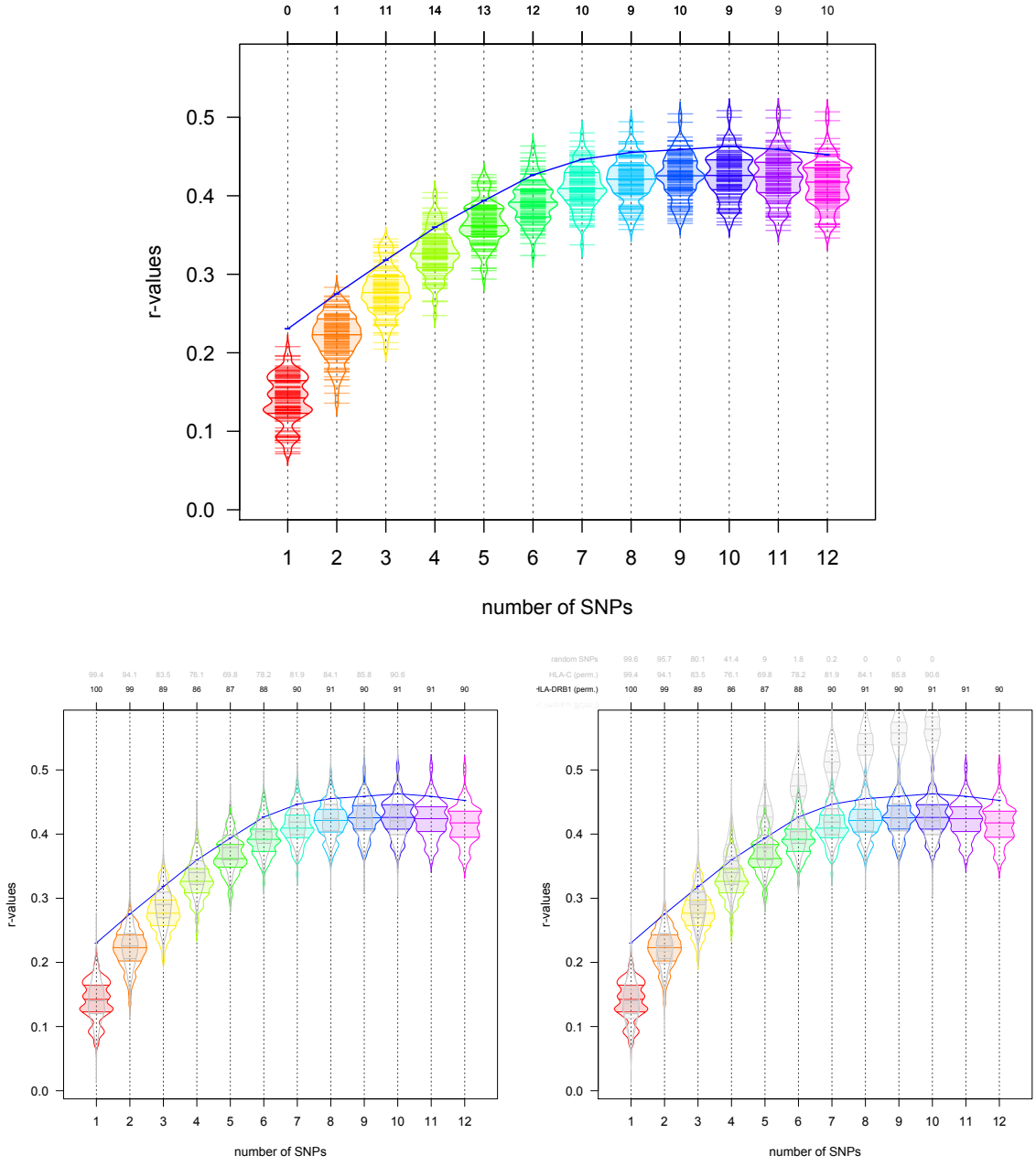

Figure 7.9: Permutation of HLA-genes and 10 random SNPs. On the top: permutation of HLA-DRB1 as violins, the blue line indicating the real data prediction. On the bottom left: permutation of HLA-C gene included as light grey violins, the bottom right also including permutation results of 10 genome-wide randomly selected SNPs. On the top margins of the plots the percentile range compared to real data.

Part III

Results

8 Procedures for Evaluation

8.1 Working process with the combined dataset

With the combined dataset, introduced in section 6.3, the following calculations were performed to create the prediction model and review the results:

- divide data into genes and expand boundaries ±200 kb, see 7.2.1
- preFilter on 0.9 correlation, see 7.1.2.2
- SVM pruning and notation of performance, see 7.2.2.2

Contrary to the idea of reducing the number of features for SVM calculations, we decided for the final evaluation of the model and to detect potentially relevant biomarkers influencing antibody production against interferon-β, to expand the gene boundaries ±200 kb. Influencing features may lie in promotor regions or close to gene boundaries. In fact, a lot of influencing features can be localized up- and down-stream a few hundred thousand base pair positions from the actual probe on the genome. For our genomic reference data, we again used the UCSC annotation database Kent et al. (2002). SNP information and position was adapted from *dbSNP* Bethesda (2005); Sherry et al. (2001). For details, see section 7.2.1 and www.genome.ucsc.edu.

8.2 Reference performance

To identify genes with significant pruning performance, a reference for comparison is needed. For this we decided to randomly select 110 SNPS genome-wide (5 SNPs per chromosome) and to once again perform a permutation. This means a SVM pruning with 10 of the 100 selected SNPs was performed. Limiting the pruning to 10 extracted SNPs per permuation kept the calculation time to a couple of days. The permutation was repeated for a total of 1 million iterations. The figure 8.1 shows the resulting reference curve of achieved performances from the permutation.

The randomly selected SNPs may yield a higher performance compared to gene-wise calculations, since they show no dependence (LD) and more variance in their genotypes. Therefore, only genes exceeding the reference can be assumed to be relevant to antibody production. To compile the significant gene selection even more strictly a modified reference curve adjusted for multiples testing was drawn: Each modified threshold value—for 1 to 10 SNPs included—results from the mean value `mean()` of the permutations result plus `qnorm()`, which indicates the distance to the 2.5 % probability of a two-sided normally distributed curve. Having a total of approximately 20.000 unique genes and calculations with 10 SNPs for each permutation the

equitation demands a distance of $2.5 \cdot 10^{-7}$ for a 5 % probability divided by 2 for each tail of the bell curve, see Fig. 6.3 on page 50. A multiplication by the standard deviation `sd()` is needed, since it is not equal to 1 in our data.

$$\text{threshold}_{\text{modified}} = \texttt{mean}(x) + \texttt{qnorm}(1 - 2.5 \cdot 10^{-7}/2) * \texttt{sd}(x) \tag{8.1}$$

This assumes a close to normal distribution within each of the violins which is then extrapolated to the desired significance level. An acceptable accordance of observed and expected performances is shown in the respective QQ-plots from 2 to 10 SNPs within the permutations, shown in figure 8.2 so we attempted this estimation.

This way, we can detect all significant genes, which exceed the reference curve performance in *one* or more times. This means that if the single SNP performance of a gene outperforms the extrapolated reference performance of one SNP, it will be selected. If the two-SNP pruning performance exceeds the reference value of the two-SNP permutations, it will also be regarded as significant. We repeat this analysis up to the maximum number of 10 SNPs. In other words: independent of the size of the model (from one to ten SNPs), a gene with at least *one* performance higher than the corresponding extrapolated reference performance will be considered as significant and selected for further interpretation.

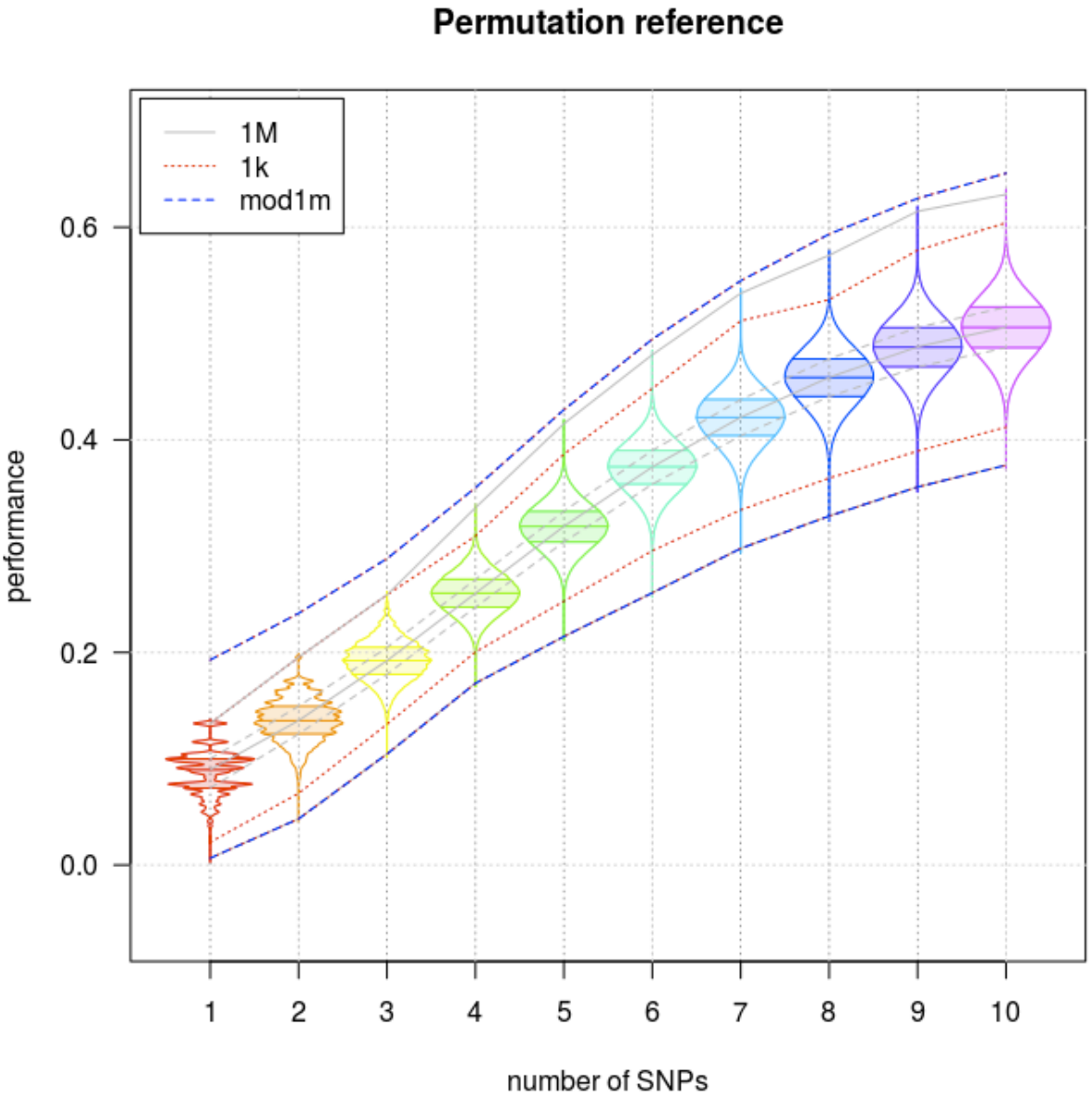

Figure 8.1: The figure shows the reference pruning performance as violins, from 1 to 10 SNPs on the x-axis and the performance value on the y-axis. The red line indicates the absolut top performance after 1000 permutations, the grey line after 1 million permutations, and the blue dotted line the modified reference curve of 1 million permutations corrected for multiple testing.

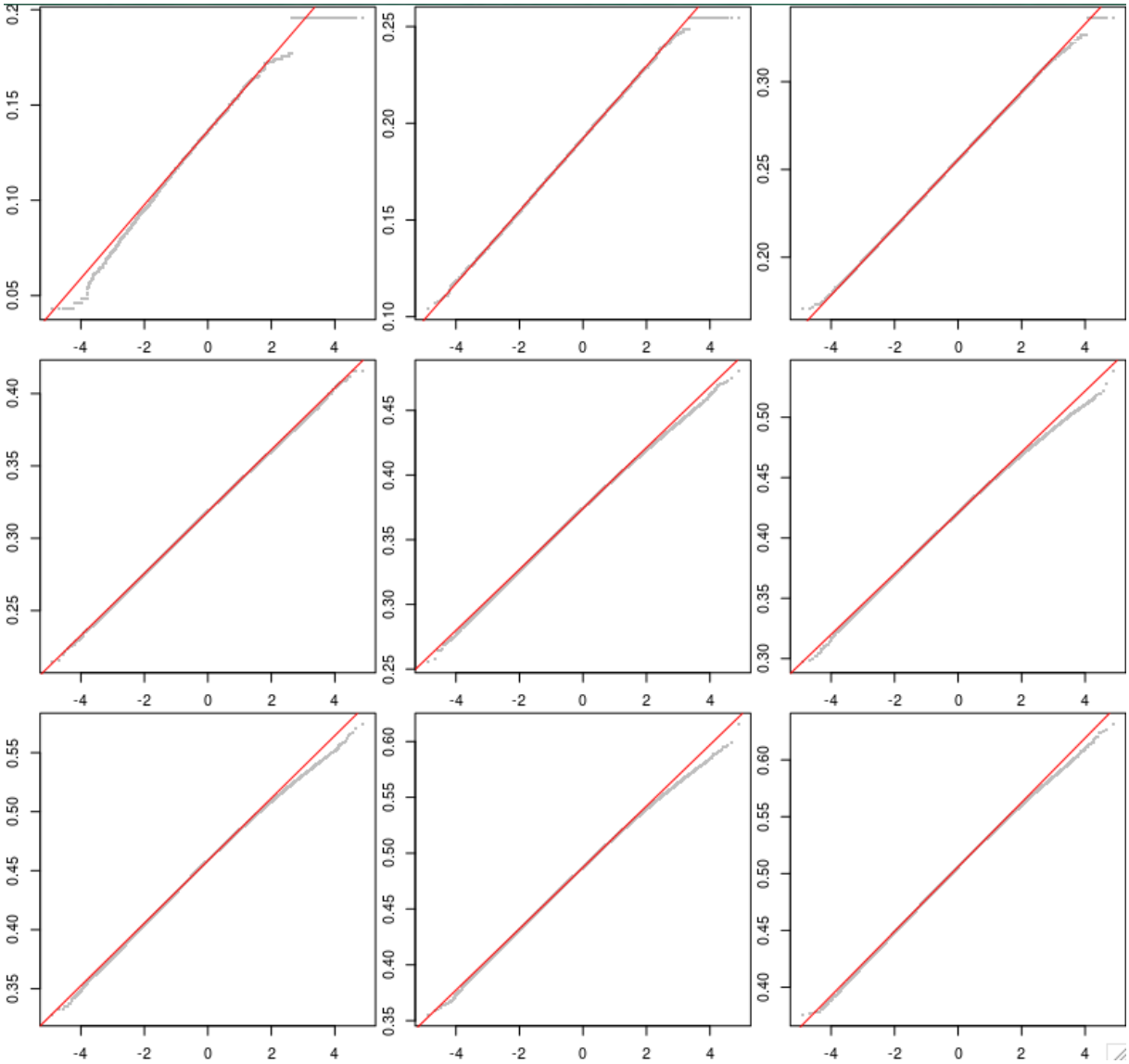

Figure 8.2: QQ plot of the expected (x-axis) and observed (y-axis) performances of the permutation. From top left indicating 2, reading horizontally to the lower right indicating a total of 10 SNPs being included to pruning calculations.

9 Results of Chromosome 6

We first evaluated the results on chromosome 6. They show that 13 genes, listed in the table 9.2, yield a significant pruning performance. Note that 7 genes are HLA genes, which are known to have influence on antibody production. Interestingly, all HLA genes, with exception of the HLA-DRB1 gene, contain the SNP rs34784936 as being the best residual SNP from the pruning calculation with a top single SNP performance of $r = 0.2$.

Correspondingly, the GWAS results of this particular SNP yield the second lowest p-value of $2.457 \cdot 10^{-8}$ out of over 6 million SNPs, as reiterated in table Table 9.1. This accordance seems to reveal a certain plausibility within our calculations.

SNP	alleles		frequency allele 1	info score	β	SE	p-value
rs34784936	G	T	0.8461	0.7448	-13.6487	2.4254	$2.457 \cdot 10^{-8}$

Table 9.1: GWAS result of the pruning top SNP rs34784936

The HCG23 and BTNL2 genes are located in short distance to the HLA genes, as shown in table Table 9.2 on the next page, and contain the SNP rs35380574, with the best single SNP performance of a comparable r-value of 0.19. They do not include the HLA pruning top SNP but upon closer inspection, although not in LD, these two top SNPs show a genotype correlation of 0.94, which explains that after preFiltering with a correlation threshold of >0.9, only one of the SNPs can be left in the genewise dataset. The HCG23 gene, which is an abbrevation for HLA complex group 23, is part of the non-protein coding region of the HLA region, as denoted by *dbSNP* Bethesda (2005); Sherry et al. (2001). However, by extending the gene boundaries, the HCG23 and the BTNL2 gene contain the same influencing SNP rs35380574. Allel variants of the BTNL2 gene have been associated with high risk for sarcoidosis, which is an autoimmune disease of unknown origin developing inflammatory granuloma Li et al. (2009); Morais et al. (2012); Wennerstroem et al. (2013). It is presumed an infectious trigger may provoke the immune system to overreact in major T-cell proliferation and damage own tissue. This hypothesis shows similarities to the etiology of multiple sclerosis and may explain the close localization of influencing genes on the genome.

Since SVMs— in contrast to other machine learning and gene analysis programs—can account for interactions, some genes may give us impressions of a higher prediction power due to SNP-interactions. For example, the DYNLT1 gene on chromosome 6, represented in the upper left corner of the figure Fig. 9.1 on page 87, shows no particulary noticeable performance unless all of the top four SNPs ranked by SVM pruning are included. A significant drop in performance can be observed when the SNP rs2919753 is excluded while pruning calculations. On the

other hand, including even more SNPs to the data can not achieve significant improvements of performance.

This means that the top ranked four SNPs rs2919753, rs9355655, rs341122 and rs3102978 show multiple SNPs interactions. Similarly, the PKHD1 and the LINC00518 genes appear interesting due to SNP interactions. While the performance of one SNP does not appear significant, the performance after adding the second SNP yields significant pruning results. See figure 9.1 to find the pruning plot for the 13 significant genes on chromosome 6. These examples indicate candidate genes in association to antibody production against interferon-β, which could not be identified with single SNP methods.

Also notable are four genes on chromosome 6, in particular DYNLT1, LINC00518, MDGA1 and PKHD1. Since they are localized outside the HLA region, they might imply independent influence to antibody production. For the gene positions see table 9.2.

gene	start position	end position
BTNL2	32362512	32374900
DYNLT1	**159057506**	**159065818**
HCG23	32358286	32361468
HLA-DQA1	32605182	32611429
HLA-DQA2	32709162	32714664
HLA-DQB2	32723874	32731330
HLA-DRA	32407618	32412826
HLA-DRB1	32546546	32557613
HLA-DRB5	32485153	32498006
HLA-DRB6	32520489	32527779
LINC00518	**10428017**	**10435055**
MDGA1	**37600283**	**37665766**
PKHD1	**51480144**	**51952423**

Table 9.2: Significant genes on chromosome 6. Highlighted are candidate genes localized outside the HLA region.

To summarize the candidate SNPs related to antibody production, we decided to execute one pruning calculation including all significant SNPs localized on chromosome 6 to one dataset. This was performed by choosing every SNP exceeding the extrapolated reference performance (SNPs marked red in Fig. 9.1 on the next page) and every higher ranked SNP in a gene's performance to include potential SNP interactions. A total of 24 SNPs were selected. To exclude high correlated SNPs a prefiltering `preFilter()` on genotype correlation >0.9 was performed. Within this step, only one SNP rs34784936, in correlation with rs35380574 as mentioned above, was excluded. Experimentally, we exchanged the two correlating SNPs. Within the genotype data of the HLA-DRA gene, we excluded the top SNP rs34784936 and replaced it with the correlated SNP rs35380574. So once more, calculations with the 23 significant SNPs on chromosome 6 were performed. Pruning computed comparable results indicating that it makes no difference which SNP to choose. Figure 9.2 shows the pruning result compared to the reference curve. Notable is the overall pruning performance of all selected 23 SNPs as it reaches an total r-value of 0.7573, as shown in figure 9.3. As metioned before in section 7.1.2.1 of this thesis, so far a single-SNP effect of only 2.6 % could be associated with the antibody titer Weber et al. (2012). The SVM calculation with 23 selected top SNPs could achieve a r-value of 0.7573 — this means a three-fold increase of prediction power could be achieved when accounting SNP interactions. The SVM prediction plot, figure 9.4, displays the impressive correlation of the measured and predicted antibody titer.

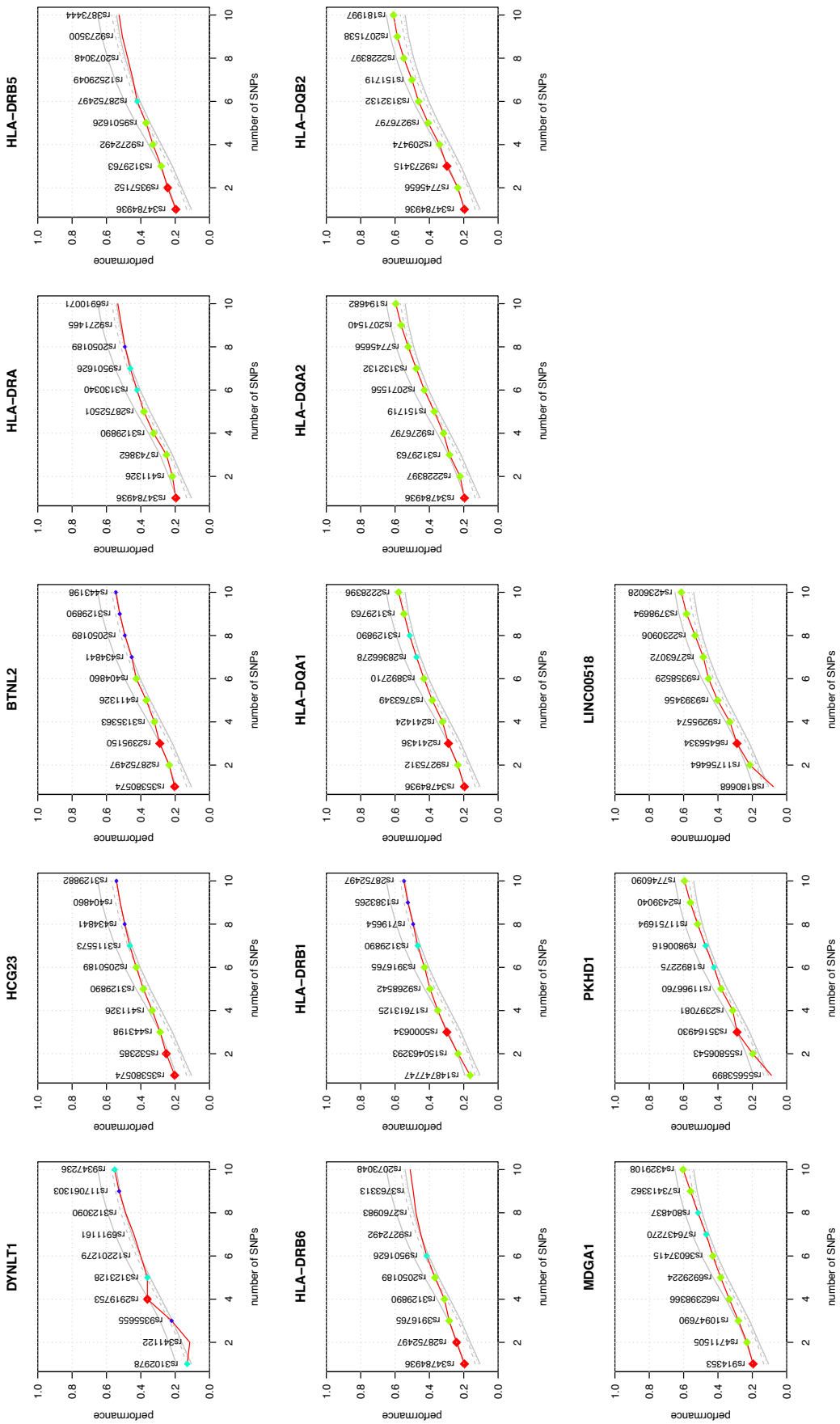

Figure 9.1: Pruning results in comparison to referance performance of top 13 genes on chromosome 6 with higher performance than permutation pruning. Red denoted SNPs indicate a performance over 100% of the extrapolated reference curve, SNPs marked green exceed 99%,SNPs marked light blue 95%, and dark blue colored SNPs indicate a performance over 90 % of the permutation results.

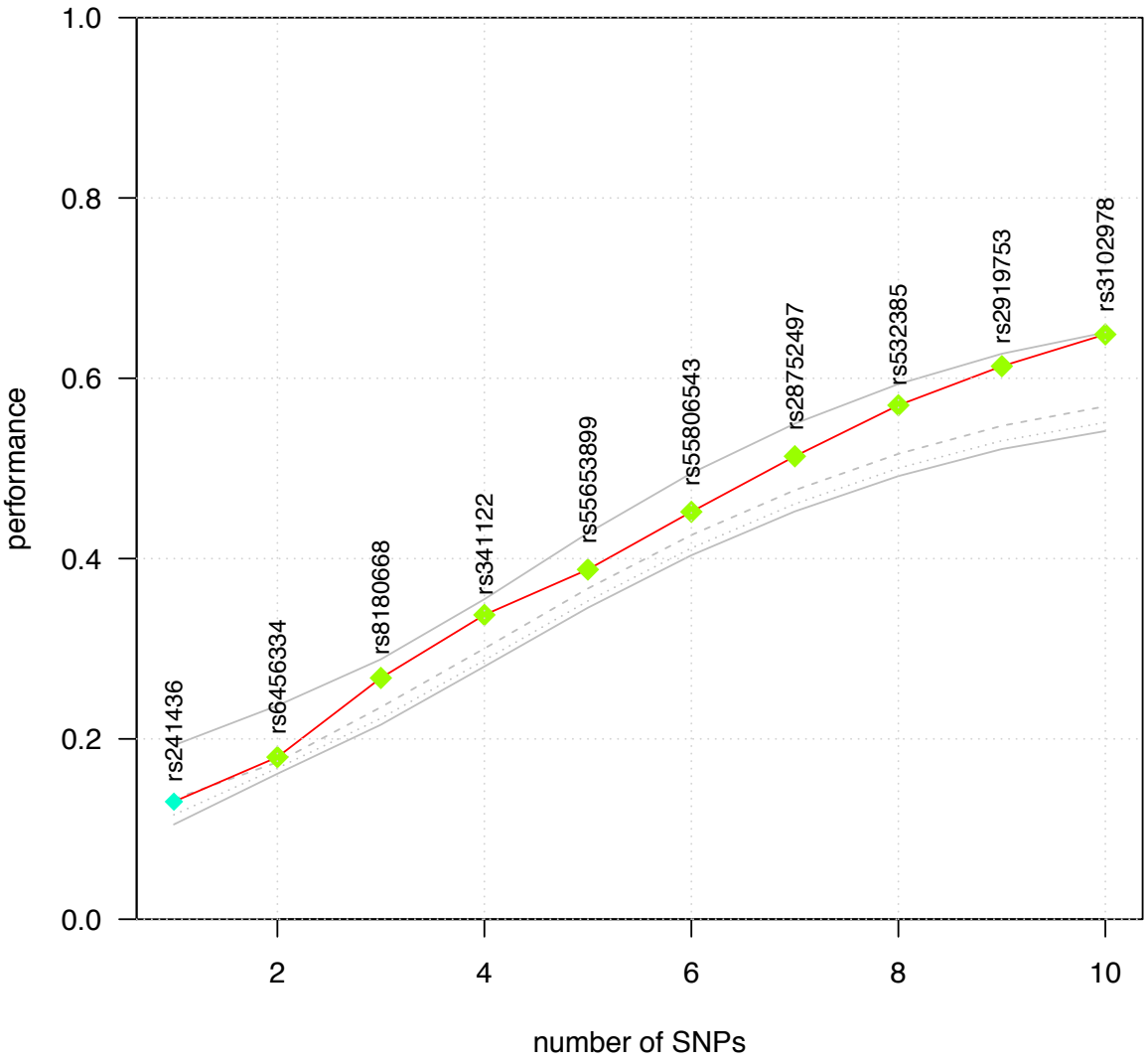

Figure 9.2: Pruning results in comparison to referance performance of combined significant SNPs on chromosome 6. Green marked SNPs exceed 99% and light blue marked SNPs outperform 95% of the reference results.

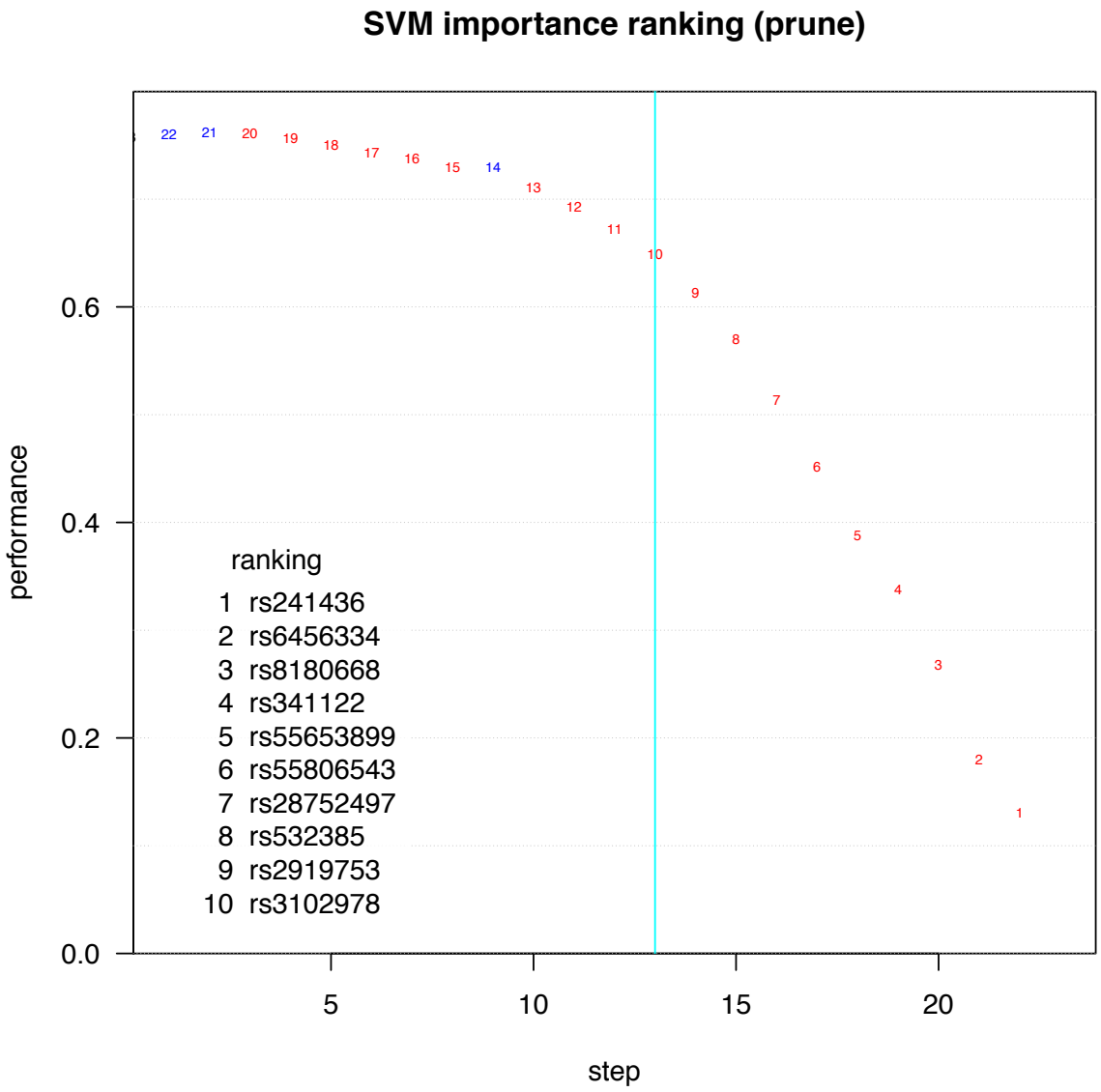

Figure 9.3: Pruning plot of summarized significant SNPs on chromosome 6. The blue vertical line indicates the 10 SNP mark, dividing the top 10 ranked SNPs to the right side of the plot. The final performance when including all 23 significant SNPs reaches an r-value of 0.7573.

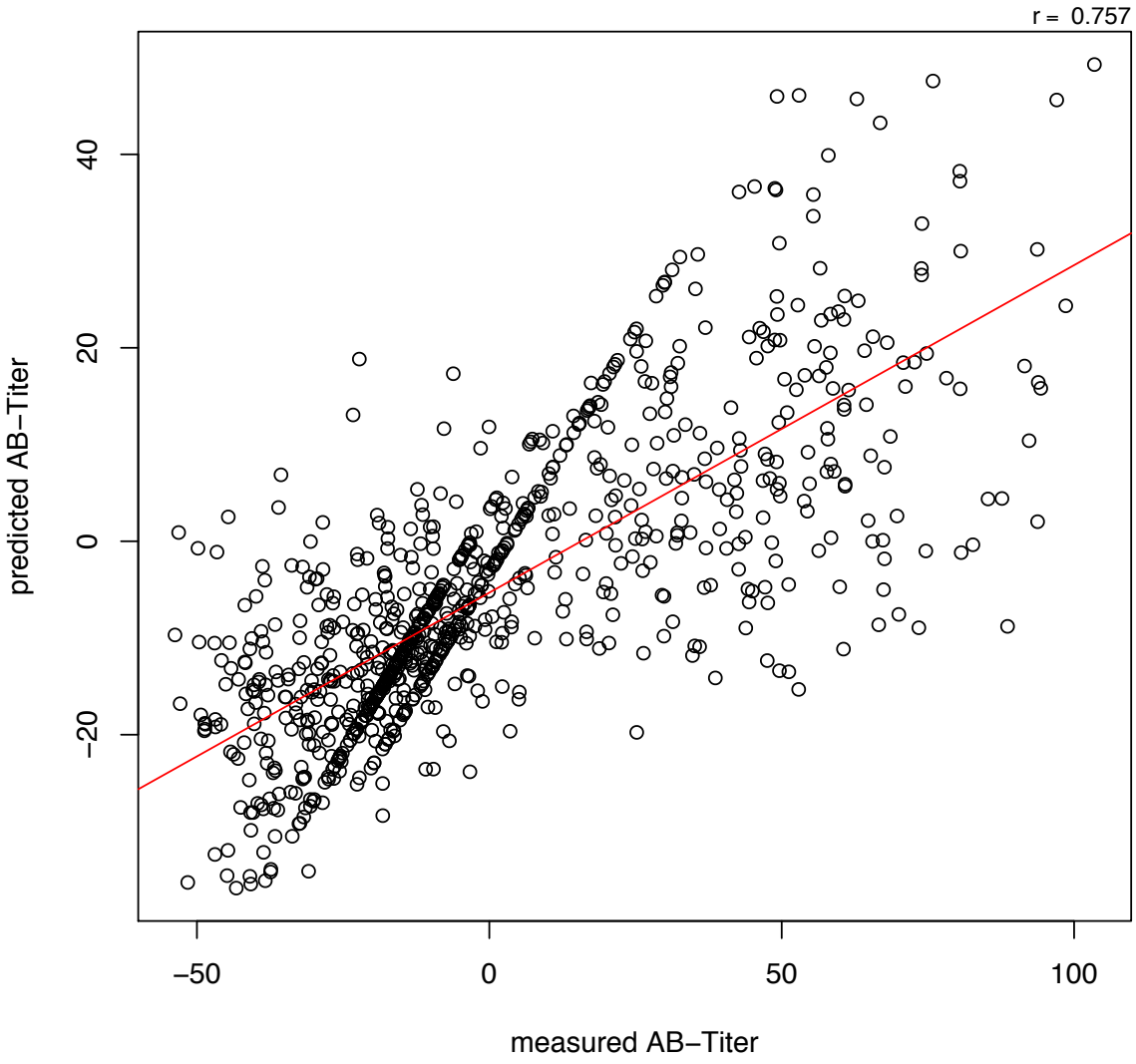

Figure 9.4: SVM prediction plot of the measured versus predicted antibody titer. The dataset contains 23 significant SNPs localized on chromosome 6 and reaches a r-value of 0.7573.

10 Genome-wide results

Extending this approach to consider all genes genome-wide, we find that a total of 78 genes exceed the extrapolated reference. In addition to the genes on chromosome 6, many other genes on various chromosomes achieve remarkable pruning performances. Figure 10.1 shows their distribution on the genome highlighted as red dots. With a total of 9 significant genes each, chromosomes 15 and 20 definitely attract attention. With the inclusion of chromosomes 1 and 19, each containing a total of 8 significant genes, these chromosomes turn out to be the most represented chromosomes within this studys results. A list of the gene names is provided in table 10.1. Furthermore, a list of the 315 significant SNPs and their base position is enclosed in the appendix.

The pruning performance plots in comparison to the reference curve for each gene are displayed in figures A.1 a to A.1 d, which can be found in the appendix.

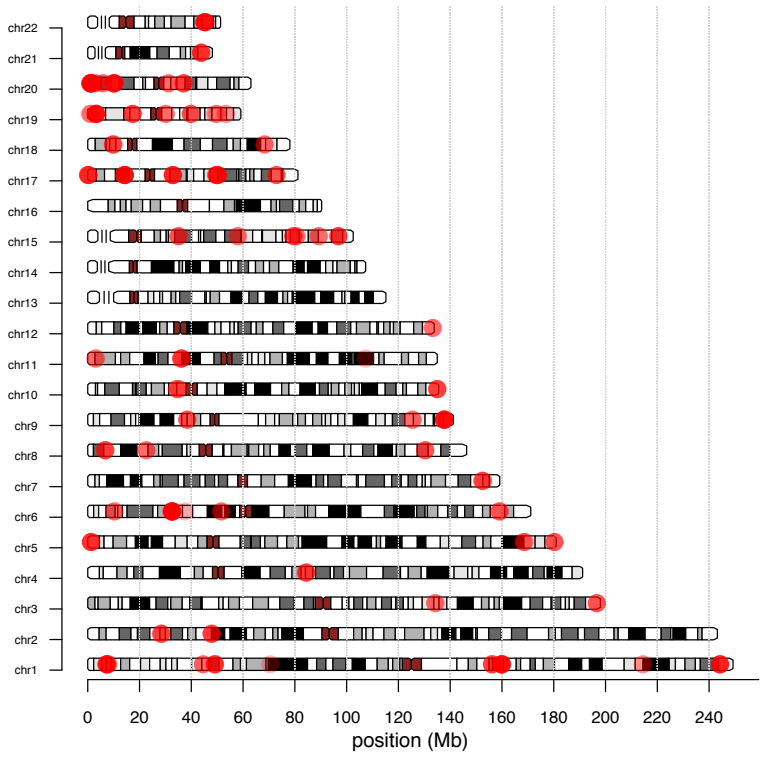

Figure 10.1: Significant genes exceeding the extrapolated referance performance are displayed with localization on the genome. The x-axis indicating the base position, y-axis indicating the chromosome.

Chromosome 1	CAMTA1	Chromosome 11	CARS
Chromosome 1	BGLAP	Chromosome 11	ALKBH8
Chromosome 1	SLAMF9	Chromosome 11	LDLRAD3
Chromosome 1	PTPN14	Chromosome 12	POLE
Chromosome 1	LOC339529	Chromosome 15	GJD2
Chromosome 1	CCDC24	Chromosome 15	LOC101928174
Chromosome 1	AGBL4	Chromosome 15	GCOM1
Chromosome 1	LRRC40	Chromosome 15	RASGRF1
Chromosome 2	BRE	Chromosome 15	ARNT2
Chromosome 2	KCNK12	Chromosome 15	MIR1179
Chromosome 3	NCBP2-AS2	Chromosome 15	MIR7-2
Chromosome 3	KY	Chromosome 15	MIR3529
Chromosome 4	MRPS18C	Chromosome 15	NR2F2-AS1
Chromosome 5	HEIH	Chromosome 17	CA10
Chromosome 5	CLPTM1L	Chromosome 17	LOC102723641
Chromosome 5	SLIT3	Chromosome 17	MGC12916
Chromosome 6	DYNLT1	Chromosome 17	C17orf97
Chromosome 6	HCG23	Chromosome 17	C17orf102
Chromosome 6	BTNL2	Chromosome 18	GTSCR1
Chromosome 6	HLA-DRA	Chromosome 18	TXNDC2
Chromosome 6	HLA-DRB5	Chromosome 19	SNAR-A6
Chromosome 6	HLA-DRB6	Chromosome 19	CELF5
Chromosome 6	HLA-DRB1	Chromosome 19	ZNF611
Chromosome 6	HLA-DQA1	Chromosome 19	ARID3A
Chromosome 6	HLA-DQA2	Chromosome 19	USHBP1
Chromosome 6	HLA-DQB2	Chromosome 19	C19orf12
Chromosome 6	MDGA1	Chromosome 19	IFNL1
Chromosome 6	PKHD1	Chromosome 19	PLEKHG2
Chromosome 6	LINC00518	Chromosome 20	TRMT6
Chromosome 7	ACTR3B	Chromosome 20	ANKEF1
Chromosome 8	BIN3-IT1	Chromosome 20	MKKS
Chromosome 8	DEFA6	Chromosome 20	PSMF1
Chromosome 8	LINC0097	Chromosome 20	LOC101929698
Chromosome 9	IGFBPL1	Chromosome 20	FKBP1A
Chromosome 9	OR1L3	Chromosome 20	SIRPB1
Chromosome 9	MIR3689A	Chromosome 20	SNORA71C
Chromosome 10	PARD3	Chromosome 20	PTPRA
Chromosome 10	MIR202HG	Chromosome 21	TMPRSS3
Chromosome 10	MIR202	Chromosome 22	ARHGAP8

Table 10.1: List of 78 genes considered significant in pruning.

For further interpretation and analysis of the 78 significant genome-wide genes, we also performed the same procedures as for chromosome 6 in section 9. All significant SNPs exceeding the absolute extrapolated reference performance and those ranked previously - to not drop possible interactions - were selected. Again, a prefiltering was performed, interestingly not excluding futher correlated SNPs except the one highly correlated SNP on chromosome 6, see chapter 9. With a total of 315 SNPs an SVM pruning was performed, see the pruning plot in figure 10.5. Figure 10.4 again shows the pruning results in comparison to the reference curve. Including all 315 variables, a performance of 0.951 could be achieved. The notable top SNP rs35380574 on chromosome 6, which achieved a remarkable single SNP performance of 0.19 within the pruning of the HCG23 and the BTNL2 gene, is consistently ranked as top single SNP of all genomewide significant SNPs. Secondly ranked is the SNP rs6064776 within the limits of the SNORA71C gene on chromosome 20. The third ranked SNP is rs11033303, which achieved top performance localized on the LDLRAD3 gene on chromosome 11.

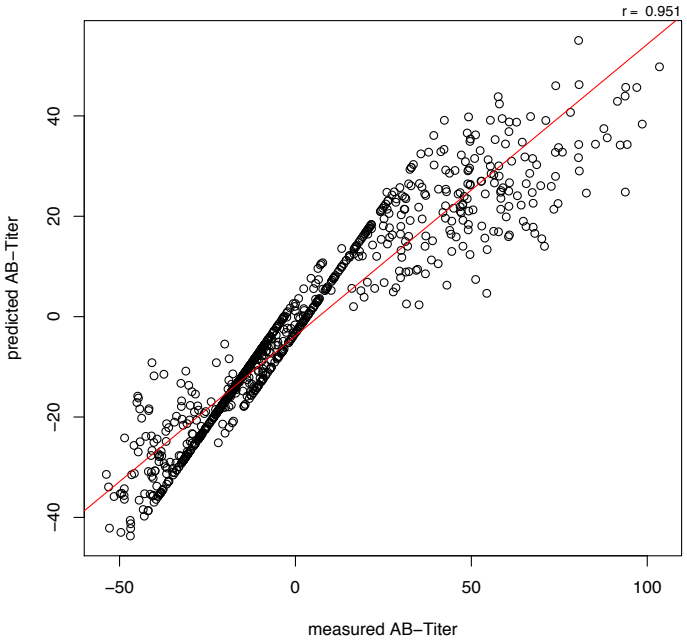

Figure 10.2: SVM prediction plot of the measured versus predicted antibody titer. The dataset contains 315 significant SNPs and reaches an absolute r-value of 0.951.

Including too many SNPs within SVM calculations may lead to overfitting - a problem of achieving artifically high correlations. In regard to not using the identical parameters in former calculations, the results in the figure Fig. 5.5 on page 36 may not be accurate to be compared. Still, we suppose the number of 315 features do reveal reliable results. Even with a clearly lower amount of features, including only 30 SNPs as Fig. 10.5 on page 96 shows, a remarkable r-value over 0.9 can be reached. With this amount of SNPs, no overfitting errors are expected. Figure 10.3 recapitulates the significant SNPs and their localization on the genome. Clearly recognizable is the HLA region localized around 30 Mb on chromosome 6. Also the chromosomes 17 to 21 contain some significant top hits, whereas chromosome 13, 14, and 16 seem to have no influence regarding the prediction of the antibody titer. The SVM prediction plot in Fig. 10.2 shows a high correlation of the measured and predicted antibody titer reaching an r-value of 0.951.

For the final model, which can be applied to predict the antibody production for each patient individually, only the 166 most relevant SNPs need to be used. This is the amount of SNPs inlcuded when the performance of the pruning calculation reaches its maximum of 0.967, as

shown in figure Fig. 10.5 on page 96. These SNPs represent the associated SNPs to antibody production in response to interferon-β medication in our study.

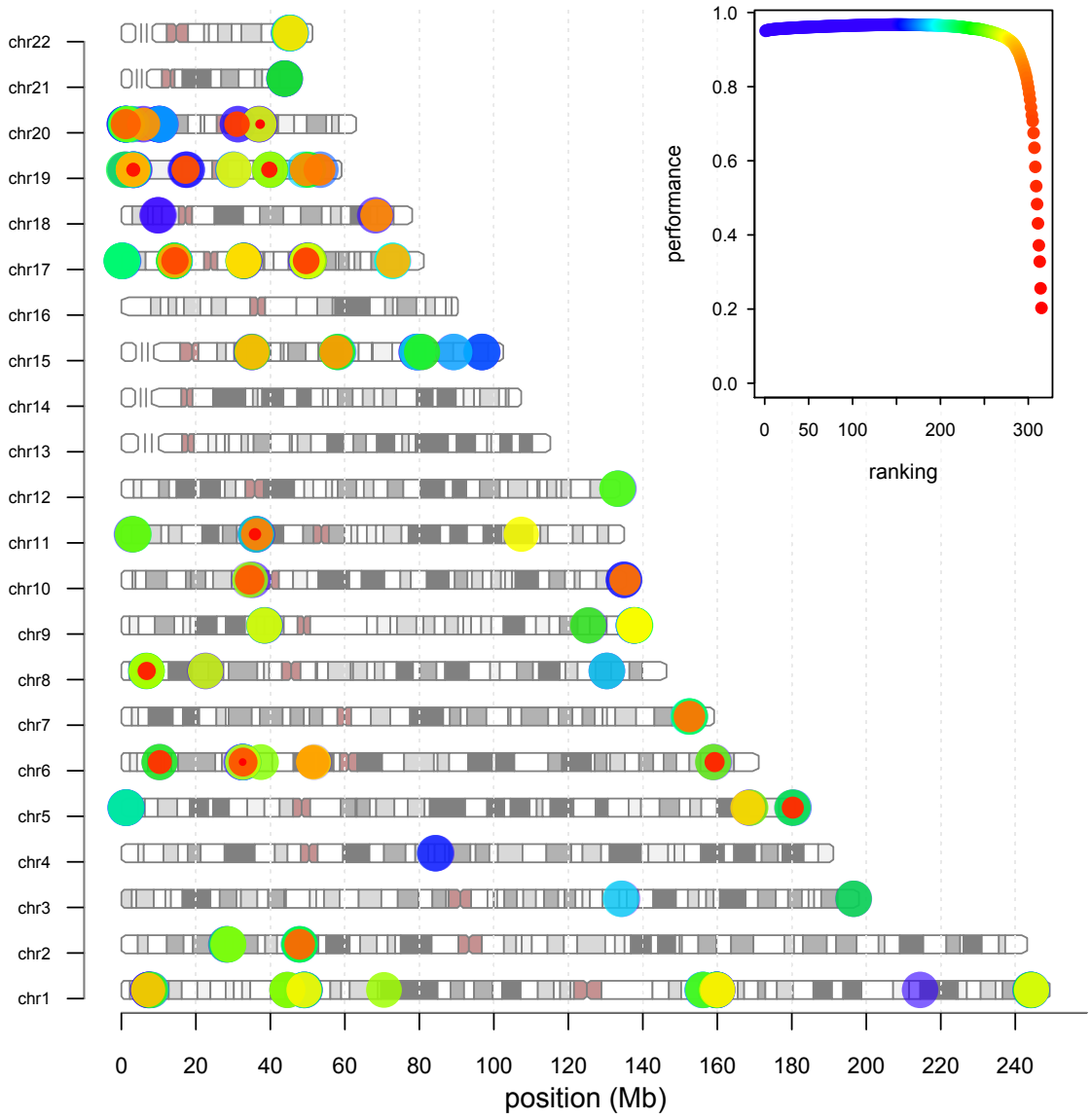

Figure 10.3: Significant SNPs displayed with localization on the genome (the x-axis indicating the base position, the y-axis indicating the chromosome) and pruning ranking - top indicating small orange over green and blue to final large violet circles. The colors are indicated in the small legend plot (which equals the pruning plot), the pointsize reflects the performance achieved when the SNP is included in the model.

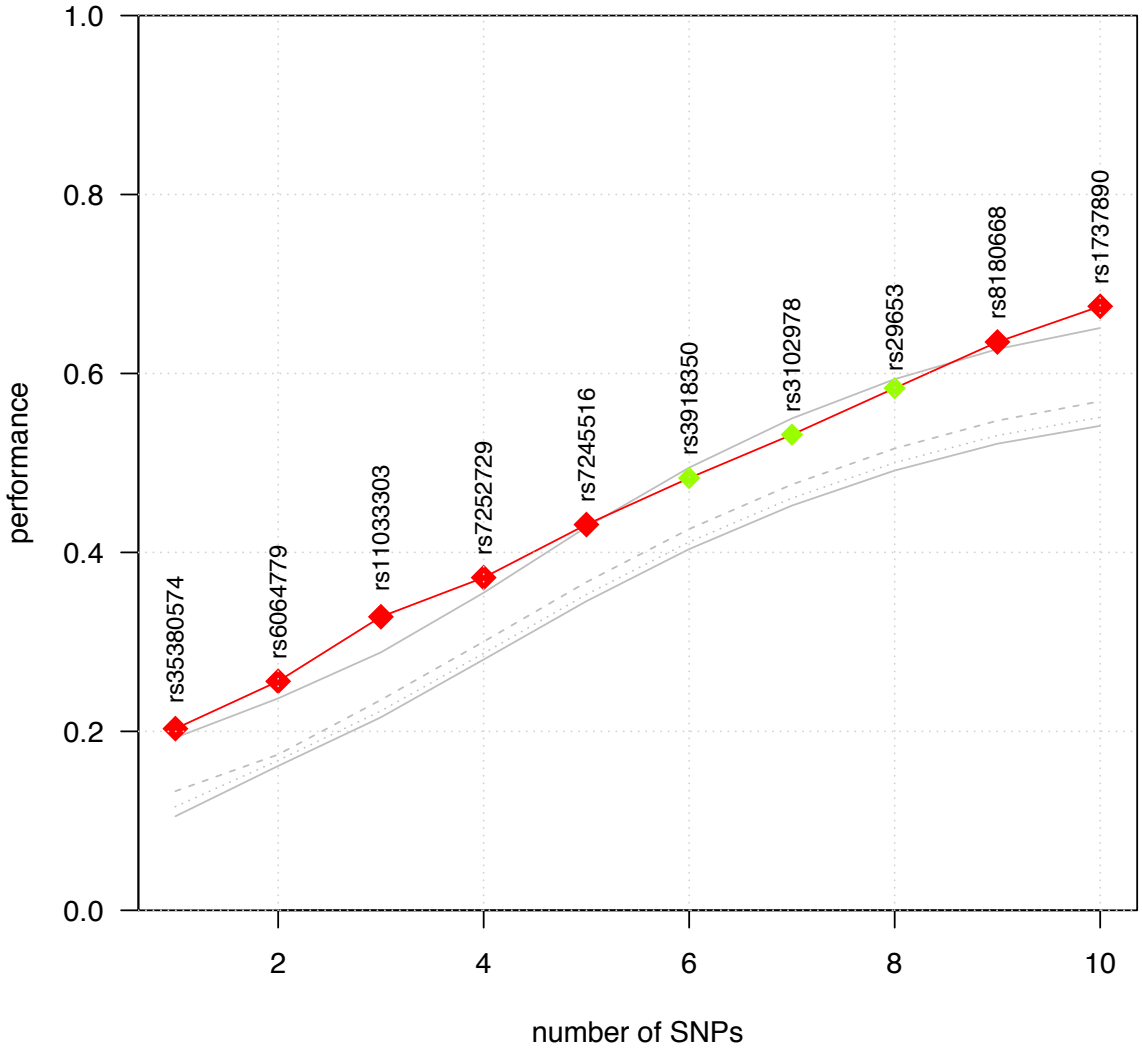

Figure 10.4: Pruning results in comparison to referance performance of all summarized top 315 genome-wide SNPs. Red denoted SNPs indicate absolute outperformance of the extrapolated reference curve, green marked SNPs exceed 99% of the reference results.

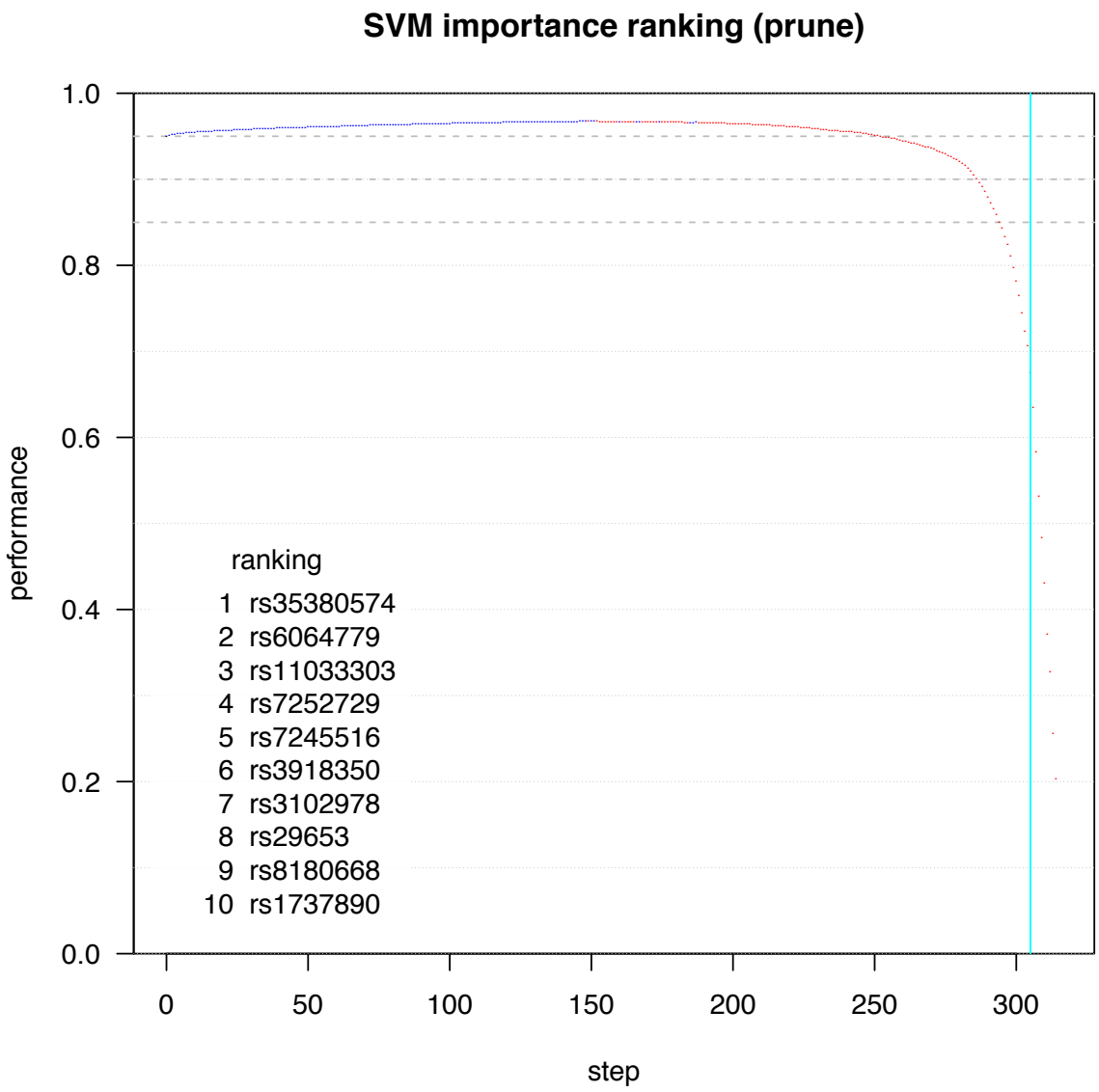

Figure 10.5: Pruning plot of all summarized top 315 genomewide SNPs. The blue vertical line indicates the 10 SNP mark, dividing the top 10 ranked SNPs to the right side of the plot. The final performance, when including all 315 SNPs, reaches an r-value of 0.951.

Part IV

Discussion

11 Review of the data

In the interim period between the creation of the SVM model and the submission of this dissertation, the data on the multiple sclerosis dataset has been modified by the Department of Neurology at the Rechts der Isar Hospital, affiliated to the Technical University of Munich. Firstly, more multiple sclerosis patients on interferon-β therapy have been recruited and patients with falsely recorded or incorrectly measured data, e. g., unreliable antibody status, have been removed from the dataset. Moreover, some patients antibody titers, and therefore their antibody status may have changed over this period of time and may have subsequently been re- evaluated. For example, a patient who initially produced binding antibodies in response to interferon-β therapy may have started to develop neutralizing antibodies at a later date, which can result in therapy failure. The SVM calculations on the prediction model have not been repeated with the most recent dataset. My model was developed and used on data available at that time. The findings are based on data collected by the Department of Neurology up to June 2014. The following section will reflect the differences between the data used for our calculations in comparison to the on-going dataset.

Further utilization of the *TUM 1 dataset* in particular should be carefully evaluated. The *TUM 1 dataset* was gathered manually and may thusly contain transmission errors. Consequently, the reliance of this data was uncertain to some extent and therefore it was removed from the updated dataset completely.

As a result of repeated measurements on the *TUM 2 dataset* consisting of 728 individuals, certain values were corrected and some were even excluded from the data. In 66 cases, a change in the antibody titer could be measured and 42 patients were eliminated. This comparison is necessary, as the data we used to create our prediction model, which would not be used again for calculations at a later stage, could have influenced the outcome of the study.

As mentioned above, due to more advanced measurement techniques when collecting data and more precise quality control, some individuals were excluded from the study after the utilization within our project. These individuals with inaccurately measured values may have biased the outcome of our calculations, which generated the prediction model. In other words we included sets of data to our calculations, which would presently no longer be used. When using such large clinical data, a certain bias concerning the patients information or the measured values has to be anticipated. Parameters such as age, sex, components of the multi dimensional scaling (MDS) analysis, the course of the disease, the EDSS score, the medication start as well as the current treatment are essential covariates of this dataset. False information, confounded values or incorrectly registered parameters would alter the dataset. Fortunately, we do not expect severe bias in our results. Although sex, age and the MDS components C1 to C5 were

included as covariates, our main focus for the final prediction model was given to the antibody titers related to the genotype. Considering the possibility that an individuals antibody titer was detected incorrectly, we anticipated the effect this would have on our SVM model. Once more, we do not expect these values to have a great influence on our calculations or bias the results. This is due to the fact that for our calculations we intentionally used individuals showing a measured antibody titer in the extreme of the distribution curve of all patients. By doing so, even falsely measured, very high titers, would not result in major changes of the dataset and miscalculations. In contrast, when calculating the median of a dataset, one extreme high value would alter the result dramatically, which is, however, not the case within this study.

Despite the mentioned defects, I believe there are no severe changes and bias to be expected in the results. Before using the dataset we repeatedly performed a precise and accurate quality control, both on the *TUM 1 dataset* and *TUM 2 dataset* as well as on the combined dataset. This way, we ensured the data was free of discernible errors or inconsistencies. Every patient's data has been carefully evaluated before being included into our calculations. Thus, the mentioned aspects need to be kept in mind.

In the future, it is favorable to gain a large homogenous dataset. It is important within a dataset that obtained values are collected, measured and obtained under same circumstances at the same point in time. This homogeneity was not provided within the data used in this study.

In our project, we examined the antibody titer against interferon-β to look for correlations regarding genotype markers. The antibodies were detected through *enzyme-linked immunosorbent assay (ELISA)*, a modern method to identify a specific substance within a sample. This technique allowed us to identify the amount of antibodies produced against interferon-β. In other words, the antibody titer can be determined. However, this method can only measure the extent to which antibodies bind to interferon-β. It cannot distinguish between binding and neutralizing antibodies. This means that no statement concerning the neutralizing effect of antibodies or in regard to the medications residual function can be made. The MxA exclusively measures the impact of interferon-β and therefore, only its concentration can determine the presence of neutralizing antibodies. By measuring antibody titers with *ELISA*, we could not differentiate between binding and neutralizing antibodies and so our results do not give any indication concerning treatment efficiency. Theoretically, a patient could have a very high titer of binding antibodies, which do not influence the medications effect, and show no neutralizing antibodies. This way, the patient would still gain therapeutic success. On the other hand, a minimal titer of mainly neutralizing binding antibodies can result in therapy failure. For this reason, it is important to emphasize that this model reveals the estimate of binding antibodies and will not predict the medications residual function. As a result of my work, one consideration can be the creation of a prediction model that only includes the titer of neutralizing antibodies, which attenuate the impact of interferon-β. It would be of great use to clinical practice.

Equally, this careful consideration is a critical topic in present-day research. Questions appear on what the medically sensible approach is when detecting binding or neutralizing

antibodies within a patients sample. A published study claimed that binding antibodies seem to appear first, followed by the production of neutralizing antibodies later in time Kivisäkk et al. (1997). This means neutralizing antibodies seem to develop after binding antibodies. This would pose the idea that the development of binding antibodies would always precede the production of neutralizing antibodies, which could give an indication as to when therapy rearrangements should be reconsidered. Yet, there would be time enough to change therapy before experiencing any treatment failure. On the downside, this would also mean a conversion to neutralizing antibodies is inevitable. Once the immune system recognizes the interferon-β as foreign, it will begin to fight back. Theoretically, it will only be a matter of time before a development of neutralizing antibodies can be measured.

At other point of view is that the production of binding and neutralizing antibodies may display two different independent immune reactions. The detection of binding antibodies would not influence therapy outcome, nor imply a subsequent production of neutralizing antibodies within time. This would also mean additional influencing factors, e. g., even different meaningful genes, leading to the development of neutralizing antibodies need to be considered.

So far, published studies denoted that the HLA-DRB1 gene seems to have an influencing effect on the antibody production to interferon-β therapy Barbosa et al. (2006); Buck et al. (2011); Buck and Hemmer (2014); Hoffmann et al. (2008); Link et al. (2014); Soelberg Sorensen (2008); Weber et al. (2012). Nevertheless, this does not necessarily implicate a neutralizing effect on the medication and therefore lead to therapy failure. On one hand, the intensity of immune response and consequently high titers of binding antibodies may induce the formation of neutralizing antibodies. On the other hand, other components, such as drug dose, length of treatment or even other undiscovered genetic markers, may be relevant in the development of neutralizing antibodies. Accordingly, a study considering the neutralizing antibodies as the given phenotype is needed. A prediction model on the neutralizing antibodies may be developed in the future and possibly result in even more precise prognosis concerning the therapy outcome.

In short, the aim of this thesis was to create a SVM prediction model. The available genotype data of multiple sclerosis patients treated with interferon-β and their corresponding antibody status provided us with the necessary factors for our calculations. Notwithstanding, the model can be reviewed carefully for further development, reconsidered for improvements as well as constantly re-evaluated with more reliable data. Furthermore, a validation of the model with a larger homogenous dataset is of great interest.

12 SVM Limitations

In our study on multiple sclerosis patients treated with interferon-β, we created prediction models to forecast antibody production against interferon-β.

Firstly, an increase of prediction power could be achieved when including more features (SNPs) to the study. This, on one hand, can display the addition of influencing single-SNP effects. Many small combined single-SNP effects will increase the overall significance of the model. On the other hand including more features can increase the prediction performance even if they are not in direct relation to antibody production. This was the case when we extended the gene boundaries by 200 kb – the climb of the performance does not necessarily indicate an inclusion of indicative SNPs for prediction. This may much more likely reveal a false positive performance when the threshold of the maximal number of included features is exceeded. In other words, regardless of the promising results some calculations can yield, a very high performance may also reflect a false positive result, which is referred to as overfitting. This is a process of calculations that includes a lot of data, which creates overly good models, when working on too many parameters. Since in this case, SVM can forecast prediction values almost perfectly with one data, it can not be implemented reliably for another dataset. In such cases, some preprocessing (excluding uninteresting, or highly correlated parameters and splitting data into suitable partitions) is required to avoid this problem.

Secondly, in contrary to the addition of single-SNP effects, SVM are able to consider possible SNP interactions, to again increase prediction accordance. A SNP interaction can be observed for example due to an excessive change of performance when adding or removing one of the interacting SNPs from the calculation. The inclusion of possible SNP interactions makes SVM a superior method compared to many other machine learning techniques.

The applied `pruning` calculations require a lot of computational capacity and calculation time, since a great amount of SVM models are created for each step. These calculations do not only provide the overall performance of the model, but also rank features on their contributional influence to the prediction model. Through initial reduction and preprocessing, as mentioned above, the disadvantage of high expenses can, again, be eluded. Once the model is created and the indicative SNPs are detected, only this small set of SNPs must be included to the prediction model for the individual patient. In this project we detected the 315 most relevant SNPs for antibody production against interferon-β. This knowledge allows for the creation of a reliable prediction model including only a small number of 166 SNPS — those that achieve the r-value of 0.967 within pruning calculation. This also means that genotype sequencing only of 166 SNPs is needed from one patient, to be taken as basis for creating the SVM model.

During the course of this thesis, we also investigated other machine learning methods to look out for advanced techniques, which may outperform SVM. For example we looked at *Random Forest* Liaw and Wiener (2002, 2015). This program did not show overfitting, as the inclusion of almost unlimited features is possible. Nevertheless, results on simulated data, which included fake highly correlated features, did not rank the indicative features in the desired top position but in the top thirty percent margin at best. Furthermore, calculations with *kernel PCA* Schölkopf et al. (1997) did not outperform SVM.

SVM reveal promising and equally importantly also reproducible results within this project. Significant genes, known to be associated with antibody production, were detected on basis of genotype data of multiple sclerosis patients. It is conceivable that the implementation of machine learning techniques will be increasingly important to help answer various medical questions in the future.

13 Discussion of the results

In this project, the SVM prediction models were created on basis of genotype information of multiple sclerosis patients treated with interferon-β. The performance was evaluated in regard of the correlation of the predicted value and the measured antibody titer against interferon-β. After data was partitioned into genes-wise subsets, the significant genes and corresponding SNPs exceeding a reference curve were selected. The results displayed a total of 13 genes on chromosome 6 which yield a significant pruning performance, as shown in Fig. 9.1 on page 87. Note that 7 genes are HLA genes, which are known to be part of the immune response pathway and have influence on antibody production. The HLA genes encode for the for the expression of immunomodulatory agents and the development of MHC class II molecules in the context of the immune response to foreign antigens. The organism response to invasive unfamiliar agents, such as bacterial or viral antigens, in T-cell proliferation and antibody production directed specifically against the attacking agent.

Amazingly, all HLA genes, with exception of the HLA-DRB1 gene, obtain the SNP rs34784936 as being the best residual SNP from the pruning calculation with a top single SNP performance of 0.2. This shows that regardless of the variation of features included within a gene-based calculation in our study, the SNP rs34784936 achieves a constant top performance. Correspondingly the GWAS results of this particular SNP yields the second lowest p-value of 2.457 x10^{-8} out of over 6 million SNPs, as listed in Table 9.1 on page 85. The results of the *GWAS* and SVM calculations assure an encouraging finding of an indicative biomarker within the prediction of therapy response to interferon-β.

Moreover, further SNPs have been associated with antibody production against interferon-β in previous studies. In particular the SNP rs9272105 localized close to the HLA-DQA1 gene, and the SNP rs4961252, localized on chromosome 8, show genome-wide significance Weber et al. (2012); Buck and Hemmer (2014). In addition, the SNP rs5743810 within the TLR6 gene on chromosome 4 does as well reveal a correlation to the production of antibodies against interferon-β in males, whereas not in females. However, note that the authors of the study state that further studies are needed to verify this proposition Enevold et al. (2010). Unfortunately, the SNPs rs5743810 and rs4961252 did not reach significant results within this study. The SNP rs9272105 did not appear within our data to shed more light on the findings.

13.1 Gene pathway analysis

The *PANTHER* (*P*rotein *AN*alysis *TH*rough *E*volutionary *R*elationships) Classification System is a web service used to analyze and classify genes and their functions Mi et al. (2016, 2013a,b);

Mi and Thomas (2009). Information about a gene's molecular function, biological process, cellular component, protein class, and pathway is provided. The available information of gene functions from *PANTHER* will be discussed in this section. Please note that again additional information of a genes localization and function are retrieved from the *Database of Single Nucleotide Polymorphisms*, dbSNP Bethesda (2005); Sherry et al. (2001).

The biological process of the candidate's genes were particularly interesting for us. We expected the significant genes to be involved in some immune process in the sequence of developing antibodies. Figure 13.1 shows a pie chart of the biological process represented by the candidate genes from pruning calculations. Note the number of 82 genes occurs since not all and on the other hand some genes where found repeatedly in the *PANTHER* database.

The red slice indicates the proportion of genes involved in immune system processes. In this category, a total of 11 genes, corresponding to 13.4 % of all included genes, were found related to the immune response. This includes, as expected, the genes within the HLA region on chromosome 6, including HLA-DRA, HLA-DRB1, HLA-DRB5, HLA-DQA1, HLA-DQA2, and HLA-DQB2. In addition, five other genes were found to contribute effects within the immune system response:

SLAMF9 is the number nine member of the SLAM family localized on chromosome 1, which are involved in B-cell mediated signaling. The gene codes for a trans membrane protein with two extracellular immunoglobulin domains involving antigen detection and one intracellular tail to communicate with other SLAM family controlled mediators. This means the coded protein represents an important member of the defending proteins of the immune system. Once a threatening or unknown antigen attaches to a B-cells receptor it responds with the production of antibodies, the production of cytokines or antigen presentation to T-cells. *PANTHER* reveals that the SLAM family even responds to interferon-γ stimuli. This is of great interest since a recent study revealed that the variation of alleles within the interferon-γ genes might also be of importance influencing interferon-β therapy outcome Enevold et al. (2010).

The LRRC40 gene, also located on chromosome 1, is primarily involved in catalytic processes such as receptor and growth factor activity. The fact that the two mentioned genes on chromosome 1 do not occur in proximity to each other and, therefore, do not contain common SNPs, indicates their independent achievement of significant pruning results.

The PLEKHG2 gene on chromosome 19 was primarily associated with various forms of leukemia and with showing influence on lymphocytic migration when overexpressed Runne and Chen (2013). Now, *PANTHER* revealed that the PLEKHG2 gene is involved in immune defense signals—in particular B-cell mediated immunity—and neurological processes. A study confirms that the genes function of activating the *Ras/MAPK* pathway through *EGFR* (epidermal growth factor receptor) leads to modification of the neural cell morphology Sato et al. (2014).

The TMPRSS3 gene on chromosome 21 found by *PANTHER* to be involved in the immune system process also shows wide-ranging involvements in published studies. Beside activity within immune response the TMPRSS3 gene shows involvements for various functions such as lipid metabolic processes and transportation, angiogenesis, blood circulation and coagulation, hormone receptor activity and apoptosis processes. The gene is part of the serine protease

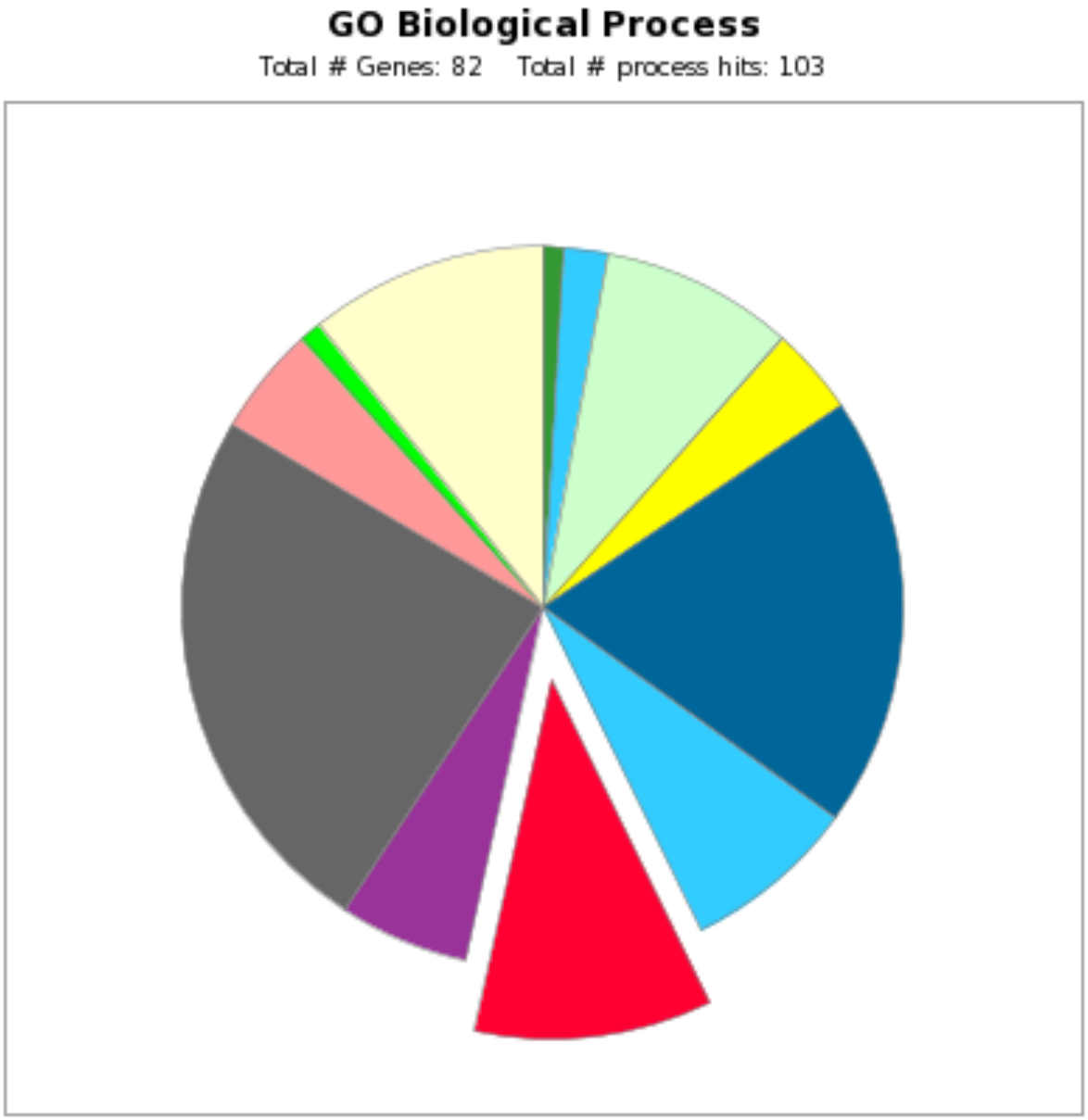

Figure 13.1: Representation of the biological processes associated with candidate genes. (apoptotic process, biological adhesion, biological regulation, cellular component organization or biogenesis, cellular process, developmental process, immune system process, localization, metabolic process, multicellular organismal process, reproduction, response to stimulus)

family and has especially been associated with autosomal-recessive congenital and childhood onset hearing loss, as it is expressed in the fetal cochlea and responsible for the inner ear development Guipponi et al. (2002). Also tumors like ovarian and pancreatic cancer have been associated Wallrapp et al. (2000).

The fifth gene identified by *PANTHER* involved in immune system process is the SLIT3 gene on chromosome 5. It is also assumed that this gene regulates cell migration, cell-cell adhesion and signaling as well as nervous system development. Interestingly a study revealed that the duplication of the SLIT3 gene may lead to major depression Glessner et al. (2010).

Subdividing the genes mentioned above involved in immune system processes, the majority of the genes indicate to be involved with antigen processing and presentation. This includes the genes localized in the HLA region of chromosome 6. Once an antigen is detected the HLA genes augment the production and expression of MCH-II molecules on the immune cells surface. MCH-II molecules present peptides of an antigen. This way cells can signal their infection and the cells apoptosis can be initiated. On the other hand, cells can prepare specific immune response directed against this antigen by developing particular antibodies to eliminate the infection. The SLAMF9 and the PLEKHG2 genes influence the immune response process. This reflects the cellular defense response, meaning in particular B-cell mediated activities are influenced. It involves cell communication such as signaling and adhesion and regulation.

Apart from the biological process, an exploration of the protein class affiliation of the significant genes also produced interesting results, as displayed in the bar chart 13.3. With a total of 7 genes, two protein classes yielded the most represented categories - classified to defense/immunity protein and enzyme modulator. Again, a coherence of the immune system involvement and the most appearing protein classes could be found. The defense/immunity protein category contains the already familiar HLA-genes as well as the SLAMF9 gene, which, as a reminder, is part of the immunoglobulin receptor family on chromosome 1.

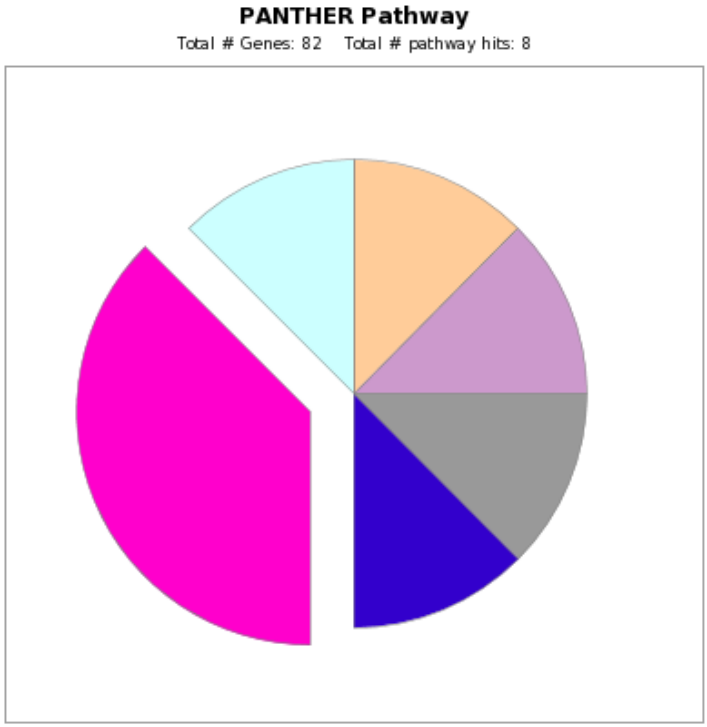

Figure 13.2: Representation of the *PANTHER Pathway Analysis* associated with candidate genes. (angiogenesis, axon guidance mediated by Slit/Robo, cytoskeletal regulation by Rho GTPase, PDGF signaling pathway, TGF-beta signaling pathway, T-cell activation)

Within the *PANTHER Pathway Analysis*, a total of 8 hits of the significant genes were detected, as shown in Fig. 13.2. In each case, a gene was found to be involved in angiogenesis, axon guidance mediated by Slit/Robo, cytoskeletal regulation by Rho GTPase, PDGF signaling pathway and TGF-beta signaling pathway. Three of the HLA genes, HLA-DRA, HLA-DQA1 and HLA-DQA2, were involved in T-cell activation. T-cells are lymphocytes, which can recognize antigens displayed on the cell surface by antigen-presenting cells. With the T-cell receptor (CD4+ or CD8+ TCR), the lymphocyte can bind and initiate immune defense.

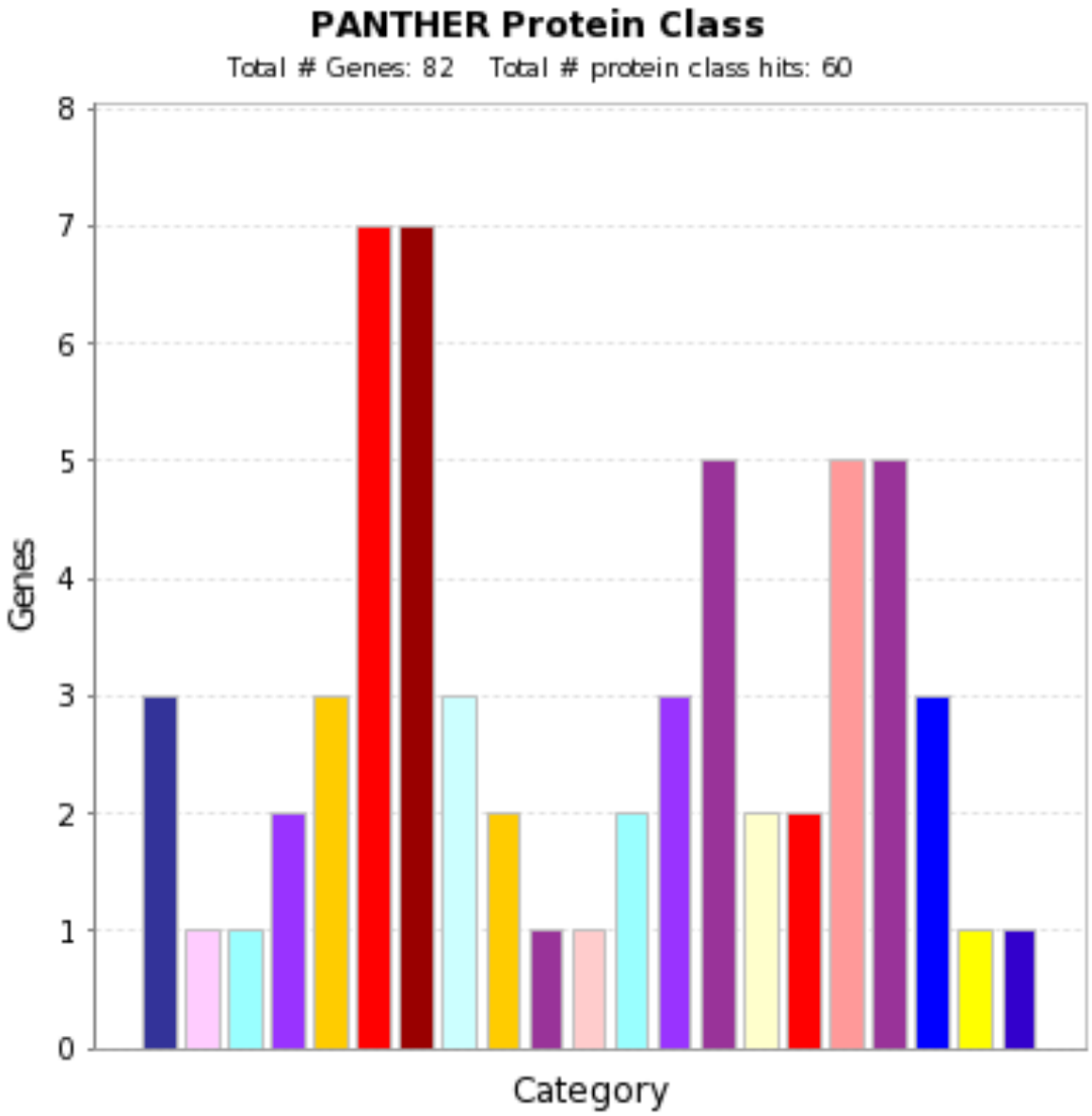

Figure 13.3: Representation of the protein class associated with candidate genes. (calcium-binding protein, cell adhesion molecule, cell junction protein, chaperone, cytoskeletal protein, defense/immunity protein, enzyme modulator, extracellular matrix protein, hydrolase, isomerase, kinase, ligase, lyase, nucleic acid binding, oxidoreductase, phosphatase, receptor, signaling molecule, transcription factor, transferase, transporter)

Finally, one more gene, TXNDC2 gene on chromosome 18 is worth mentioning. As the genes function shows oxidoreductase activity, *PANTHER* indicates a correlation to stress response. Since stress is a risk factor for developing multiple sclerosis and relapses, it may also imply an augmented risk for the development of antibodies against medication.

13.2 Consideration of additional factors

Up until this point, the discrepancy of tolerance and immunogenicity of interferon-β in patients affected with multiple sclerosis has not yet been fully understood. Although recent studies suggest an association between HLA genes and the production of antibodies, the immune process pathway still remains fairly unexplored. Previous studies revealed that a correlation between the HLA alleles HLA-DRB1*0401, HLA-DRB1*0408 and HLA-DRB1*1601 and the production of antibodies to interferon-β is assumed Hoffmann et al. (2008); Buck et al. (2011). Also, the allele HLA-DRB1*0701 especially when linked to HLA-DQA1*0201 showed an influence on the formation of antibodies Barbosa et al. (2006). Furthermore, a recent study also investigated a specific genetic predisposition for the development of neutralizing antibodies dependent of the interferon-β-preparation given Link et al. (2014). The results showed that the allele HLA-DRB1*15 causes an enhanced risk to develope neutralizing antibodies against Interferon-β_{1a} applied both subcutaneously and intramuscularly. Moreover, the well-known allele HLA-DRB1*04 revealed a high risk of antibody production especially against Interferon-β_{1b} preparations. In this study, only patients developing neutralizing antibodies, apart from those developing binding antibodies, were included. The findings of this study may be explained by a difference in binding affinity of HLA molecules to Interferon-β_{1a} and Interferon-β_{1b}. Summarizing this study indicates a genetic predisposition to immunogenicity depending on interferon-β preparation given, certainly further studies are needed.

Despite these imposing research findings, a recent study found that no direct association between the HLA-DRB1, HLA-DQA1 or HLA-DQB genes and a response to interferon-β therapy could be proven. This means that the previous assumptions cannot yet be verified Comabella et al. (2009). More and more research projects are initiated to delve deeply into exploring the immunogenicity of protein therapeutics, such as interferon-β.

Beside genetic influence within the immune response to interferon-β therapy, possible additional factors need to be carefully considered. Immunostimulatory agents such as *lipopolysaccharide (LPS)*, *lipopeptides* or other substances added during interferon-β preparation, specifically appear to influence the immune system and result in stimulating antibody production in defense to given medication, as suggested by a recent study Enevold et al. (2010). This impression can be enhanced when considering the fact that the TLR6 gene, which carries the associated SNP to antibody production rs5743810 on chromosome 4, is known to have a elementary part in pathogen recognition processes Enevold et al. (2010). This means that a residual amount of bacterial lipoprotein or other bacterial traces may appear within a protein drug and possibly trigger immunogenicity. In other words, residual substances within interferon-β preparation may lead to a boost of the immune system and induce the formation of neutralizing antibodies. Most of all, treatment with *Betaferon* seems to provoke the development of antibodies. *Betaferon* is a Interferon-β_{1b} obtained from *E. coli* bacteria or produced synthetically. Note the accordance of the thoroughly large percentage of patients within our data developing antibodies when treated with *Betaferon* (48,5 % of patients within the *TUM 1 dataset* and 40,9 %

of patients within the *TUM 2 dataset*) in agreement with previous studies. This may indicate a higher risk in producing antibodies when the protein drug is no human product. Moreover, we have to bear in mind that residual agents in protein drugs can also lead to dangerous side effects, cause allergic reactions or even an anaphylactic shock when not tolerated. Thus, there is a great focus on improving the purification process within drug manufacturing to provide safer protein drugs in the future.

What is also remarkable is the extremly low amount of patients that developed antibodies after being treated with *Avonex* in the *TUM 2 dataset* (only 9,6 %). *Avonex* is a Interferon-β_{1a} obtained from mammalian cells and therefore consists of the identical amino acid sequences to human interferon-β. *Betaferon* differs in the initial methionine as well as two other exchanged amino acid sequences, which can be difference enough to display a potential immunomodulatory factor in inducing the formation of neutralizing antibodies. *Avonex* is the only interferon-β applied intramuscularly once a week. *Betaferon*, *Rebif 44* as well as *Rebif 22* are applied subcutaneously. This low incident of antibody production within the group of patients treated with *Avonex* may also indicate a different absorbance of medication and reveal a protective effect on antibody production. Furthermore, the less frequent application may play a participating role in better tolerance of the medication.

As mentioned above, *Betaferon* is applied subcutaneously and represents the group of interferon-β preparation with the highest percentage of antibody production at the same time. A recent study revealed this statistic may be additionally explained due to the fact that the skin tissue contains a high amount of dendritic cells, which are responsible for detecting external agents. The dendritic cells might be activated through frequently injected interferon-β and initiate a immune response Link et al. (2014).

The discrepancy of antibody production within various interferon-β preparations and application methods was also denoted in further studies Barbosa et al. (2006); Buck et al. (2011); Hoffmann et al. (2008); Link et al. (2014). Note that not all of these findings accord to those detected within the *TUM 1 dataset*, which verifies once more as mentioned above that this dataset needs to be reviewed carefully before further utilization (for details see chapter 6).

Summarizing various factors such as drug manufacturing, dose application and frequency of admission as well as length of treatment need to be kept in mind since they influence the balance between protein drug tolerance and immunogenicity. Even the possibility of a missing antibody response due to a compromised immune system needs to be considered. The progress in developing less immunogenic drugs is essential, as is exploring the impact of genetic biomarkers. The prediction model created within the scope of this thesis can be a supplementary screening strategy in clinical practice to forecast therapy outcome.

14 Conclusions and Outlook

In this thesis, we investigated possible approaches to predict therapy response for MS patients to be treated with interferon-β. The correlation of genotype information and measured antibody titer against interferon-β yielded a first lead towards improving multiple sclerosis therapy. In this study, we created prediction models on basis of gene-wise datasets and selected indicative SNPs, which raised the prediction power. Once established, our SVM model only requires a small group of selected biomarkers to create and predict antibody production against interferon-β medication for each patient individually. In other words, only the indicative biomarkers, which we found to yield high prediction power, need to be regarded in further considerations. Our goal in creating such a prediction model for therapy response and subsequently the possibility of adjusting prescribed medication for each individual patient was achieved and can be applied in future medicine. This method is a powerful approach, which can avoid ineffective medication. Regardless of age, condition or course of disease, SVMs can be applied when genotype information is associated with a measurable phenotype. In the future, this technique can hopefully be generalized and applied to other questions where complex data needs to be filtered and the relevant subset of parameters extracted in order to answer specific medical questions. Applications can be envisioned for a classification into, e. g., disease subtypes, responder/non-responder cases, or regression approaches towards treatment response. The application of the SVM prediction models brings us one step closer to the favorable personalized medicine for the future.

Part V

Appendix

A Complete results for whole-genome analysis

A.1 List of 315 significant SNPs

SNP	chromosome	position	SNP	chromosome	position
rs1008298	1	7245922	rs10201794	2	28668724
rs10218507	1	49017890	rs1035872	19	17546855
rs1038973	1	244240283	rs10518915	15	57684825
rs10745297	10	134881862	rs10752633	1	159896294
rs10760244	9	125385408	rs10775365	17	72780491
rs10776668	10	134862552	rs10803182	1	244357987
rs10814685	9	38381639	rs10814689	9	38406340
rs10858286	9	137730384	rs10864296	1	7403874
rs10871511	17	126701	rs10908491	1	156208230
rs10927090	1	244073630	rs10927102	1	244174161
rs11025111	11	3053234	rs11033303	11	35871266
rs11080383	18	10076750	rs11084135	19	30214277
rs11087572	20	3212741	rs11101624	10	135003852
rs11120840	1	7156082	rs11120889	1	7324323
rs11120997	1	7720924	rs111283115	15	89344863
rs11147132	12	133451540	rs111930026	20	2754511
rs11205496	1	48969477	rs112170921	9	125628835
rs1125750	17	49902734	rs113178069	15	35087741
rs113975272	3	196491701	rs1149339	1	7536743
rs11587021	1	159825916	rs11638277	15	96760138
rs11649737	17	14237195	rs11661004	18	68285549
rs11668388	19	1098809	rs11679268	2	47873170
rs116793	20	9971941	rs11682666	2	47838155
rs11698234	20	1025258	rs11702374	21	43829758
rs11702755	21	43900692	rs11737901	5	1430616
rs11750211	5	1183560	rs11756464	6	10240443
rs118183060	17	49794964	rs11855058	15	79310544

Table A.1 a: List of significant SNPs genome-wide after SVM pruning, part 1

SNP	chromosome	position	SNP	chromosome	position
rs11857451	15	89321584	rs11871882	17	14129359
rs11876181	18	68499645	rs12439411	15	80525919
rs12440569	15	96623769	rs12450662	17	72487
rs12457955	18	68128269	rs12459085	19	17567428
rs12548614	8	22560438	rs12600962	17	32895461
rs12766409	10	134969638	rs12806492	11	107352012
rs12904245	15	79515988	rs12945409	17	72796667
rs13102255	4	84524957	rs13157838	5	168896501
rs13161677	5	168622840	rs13171076	5	1120954
rs132021	22	45425233	rs1386267	17	50323521
rs1411271	9	125315778	rs143127860	2	28335060
rs143148060	1	44546452	rs147990233	19	39952475
rs148224267	1	214460766	rs1484405	1	49104821
rs148476860	10	34550696	rs148579485	2	28289041
rs148747747	6	32575078	rs150463293	6	32755842
rs150938005	8	130338309	rs1541204	20	36918992
rs1541379	17	49910764	rs1565824	1	244225402
rs163184	11	2847069	rs1660364	1	214405605
rs16959793	15	35071718	rs1737890	20	31042595
rs181311005	2	47660259	rs182632445	4	84570345
rs1865094	19	39976969	rs1877674	8	22606292
rs191795474	1	156131050	rs1966755	15	58200847
rs2038180	20	5766395	rs2185214	1	159976751
rs2191055	17	32833531	rs2232003	19	49621964
rs2235954	20	2644975	rs225430	21	43801712
rs225439	21	43731405	rs2256258	20	1518586
rs2269546	22	45254717	rs2281118	22	45002548
rs2299845	18	9807564	rs2302063	19	3150418
rs2317651	20	1414025	rs2395150	6	32326045
rs241436	6	32797876	rs2436206	8	130361779
rs2493215	1	8007716	rs2500499	1	244271051
rs2582848	10	34453743	rs27047	5	1412251
rs2735946	5	1300429	rs2741665	8	6806289
rs2741702	8	6774720	rs2840595	1	159884109
rs28451948	9	137542038	rs28482886	20	31291483
rs28528230	17	32702452	rs28536730	20	31301954
rs28698282	9	137876465	rs28752497	6	32568481
rs2912095	8	6633130	rs2919753	6	159094468
rs2951847	8	6764166	rs29653	5	180266951
rs2993129	9	38563171	rs307808	5	180069124
rs308049	19	3116725	rs3102978	6	159263542
rs310679	19	3150681	rs3109676	9	137631198
rs3123025	1	156267004	rs312926	19	3275315
rs319959	1	49145514	rs341122	6	158864800

Table A.1 b: List of significant SNPs genome-wide after SVM pruning, part 2

SNP	chromosome	position	SNP	chromosome	position
rs34436555	11	36171911	rs34538402	20	37009728
rs34636157	7	152512629	rs34784936	6	32559648
rs34866848	15	89058528	rs35164930	6	51730580
rs35244005	17	286825	rs35380574	6	32560051
rs36265	19	39862881	rs363018	20	10236231
rs3760965	19	3030052	rs3780889	10	35299715
rs3795138	20	1292606	rs3811159	9	137688657
rs3818331	20	1281685	rs3842947	15	80608149
rs3918350	8	6786781	rs3922644	9	137593392
rs4041594	17	32766923	rs4054648	17	14343290
rs424694	21	43789772	rs4254288	15	79234123
rs429034	11	36413934	rs441327	2	47982572
rs442262	20	1207138	rs4433388	1	159729337
rs445208	19	3054699	rs4512645	1	159729047
rs4516708	4	84219635	rs451778	19	3119406
rs4596720	9	137587628	rs4693608	4	84241357
rs4701016	5	180458539	rs4726173	7	152261079
rs4726203	7	152472045	rs4778791	15	80724551
rs4788857	17	73018906	rs4789112	17	72943929
rs4791571	17	14215076	rs4792477	17	14275086
rs4795950	17	32932859	rs4807478	19	3493601
rs481068	3	196720038	rs4842174	9	137722311
rs4867890	5	168185423	rs4872016	8	22652045
rs4883632	12	133198021	rs4890190	17	62820
rs4926749	1	49203349	rs4934506	10	34236915
rs4968145	17	455814	rs4975622	5	1232666
rs4992762	12	133140352	rs5000634	6	32663564
rs5024475	1	159946733	rs509880	20	897095
rs530652	20	930560	rs532385	6	32195359
rs55653899	6	51630989	rs55806543	6	51381205
rs55903142	17	160514	rs56003400	1	160112660
rs5741814	20	36979049	rs5766045	22	45202832
rs57990176	17	14377998	rs58669357	19	53185582
rs59194105	19	30116512	rs59966862	15	57928978
rs6006853	22	45115904	rs6006915	22	45432447
rs6007309	22	45223887	rs6039649	20	9983300
rs6039732	20	10121178	rs6039885	20	10351696
rs6042009	20	1382965	rs6053821	20	5984173
rs6056276	20	946366	rs6064323	20	36914386
rs6064779	20	37252342	rs6074096	20	10085320
rs6074148	20	1098486	rs6077755	20	1089961
rs6078095	20	1171788	rs6079606	20	1544162
rs608684	9	137933064	rs6087052	20	9879166
rs6107718	20	5911963	rs6109816	20	1385863

Table A.1 c: List of significant SNPs genome-wide after SVM pruning, part 3

SNP	chromosome	position	SNP	chromosome	position
rs61226110	8	130390408	rs61247522	1	160068629
rs6133850	20	10286150	rs61376711	18	9685980
rs62006070	15	34952513	rs62054200	17	14187847
rs62081250	18	9840972	rs62185615	20	10070110
rs62269488	3	134131069	rs6427299	1	156027550
rs6456334	6	10259818	rs6502354	17	14312297
rs6504752	17	49993999	rs6510919	19	728416
rs6514392	20	1390247	rs6547816	2	28169597
rs6565719	17	215798	rs6669634	1	159855986
rs6669999	1	7263923	rs6670333	1	156153535
rs6674401	1	70499373	rs6686288	1	7129505
rs66938508	19	17264097	rs67251149	3	196603520
rs67594359	19	1168467	rs67792753	19	53398266
rs6795013	3	134139460	rs6831788	4	84282828
rs6966090	7	152675764	rs7040397	9	38392256
rs707475	1	7917076	rs7107975	11	36071441
rs7124550	11	36145810	rs71484072	10	135230128
rs7164134	15	79496143	rs7169894	15	96822001
rs7180365	15	96945057	rs718202	19	3374702
rs7211270	17	49515133	rs7245516	19	39715209
rs7246865	19	17219105	rs7252729	19	3182501
rs72848265	11	3251845	rs73029471	19	30041862
rs73050931	19	49065791	rs73165329	22	45454171
rs7412689	1	44537663	rs742196	22	44974767
rs4867890	5	168185423	rs4872016	8	22652045
rs745136	15	96899827	rs7508601	19	50151686
rs7535810	1	244358308	rs75430132	9	137784552
rs758403	17	32956286	rs7603494	2	47904096
rs76125718	1	44593335	rs763213	22	45075738
rs7708552	5	180195645	rs7735612	5	168509732
rs7745656	6	32680970	rs7832303	8	130270531
rs78906436	1	49038072	rs7929961	11	36388162
rs79579224	7	152742530	rs79830201	20	10497034
rs8075272	17	33042432	rs8099879	19	17230287
rs8119636	20	1162370	rs8180668	6	10316668
rs8182303	17	49526882	rs875789	2	28700759
rs886862	17	14215436	rs893129	15	35093454
rs905449	17	50342947	rs914353	6	37474875
rs9273415	6	32627310	rs9275312	6	32665728
rs9355655	6	158901809	rs9357152	6	32664960
rs9675446	18	9703304	rs9787643	10	35267810
rs9862761	3	134502697	rs9867805	3	196685436
rs9898067	17	50411839	rs9899546	17	14256093
rs9899744	17	49613095	rs9978405	21	43891087

Table A.1 d: List of significant SNPs genome-wide after SVM pruning, part 4

A.2 Performance plots of significant genes

The following pages show the performance plots for all 78 genes where at least one SNP exceeds the reference performance. These genes represent the significant results associated with antibody production against interferon-β in our study.

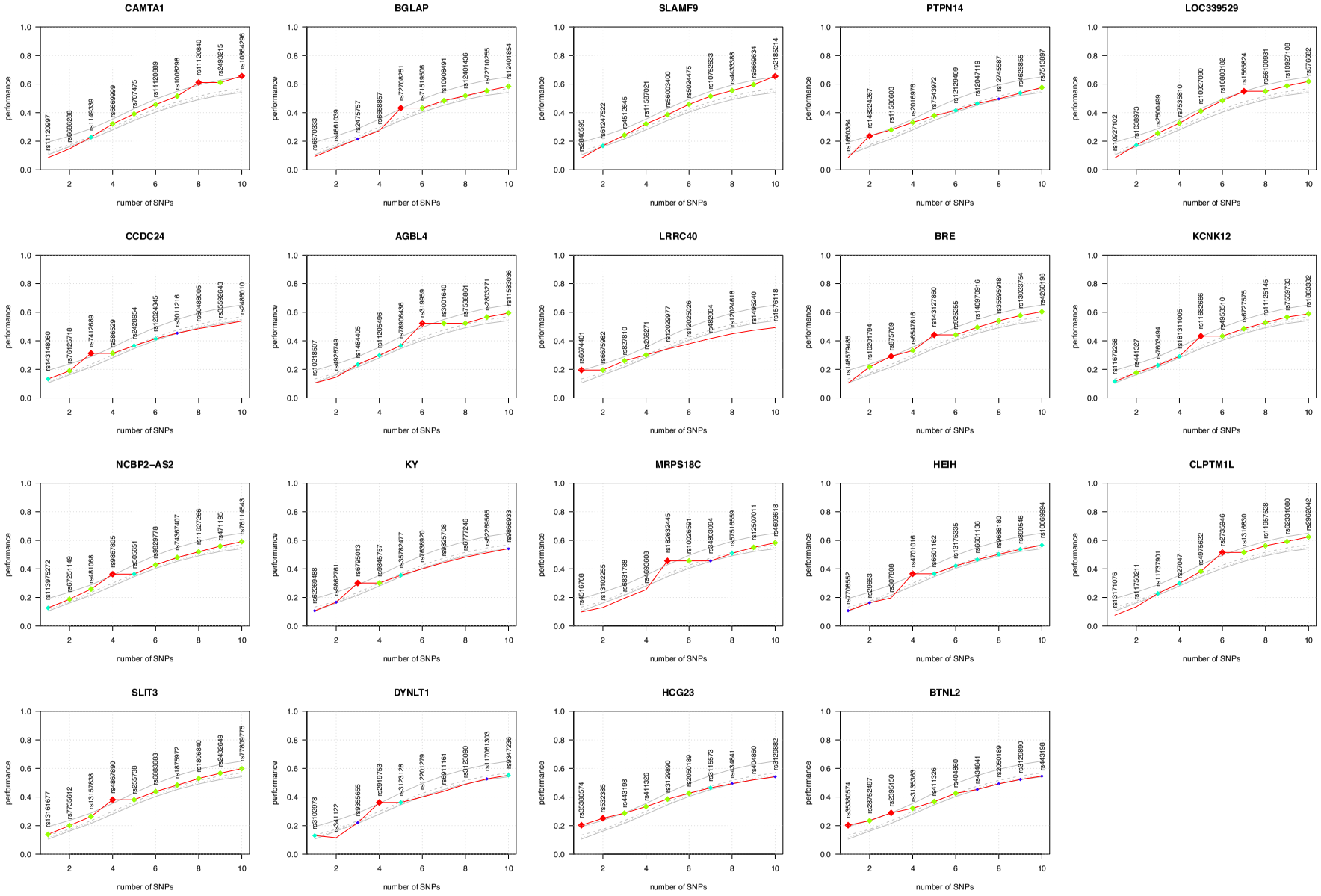

Figure A.1 a: Top genes - chromosome 1 to 6. Red denoted SNPs (◆) indicate a performance over 100 % of the extrapolated reference curve, SNPs marked green (◆) exceed 99 %, SNPs marked light blue (◆) 95 %, and dark blue colored SNPs (◆) indicate a performance over 90 % of the permutation results. SNPs below 90 % are not marked.

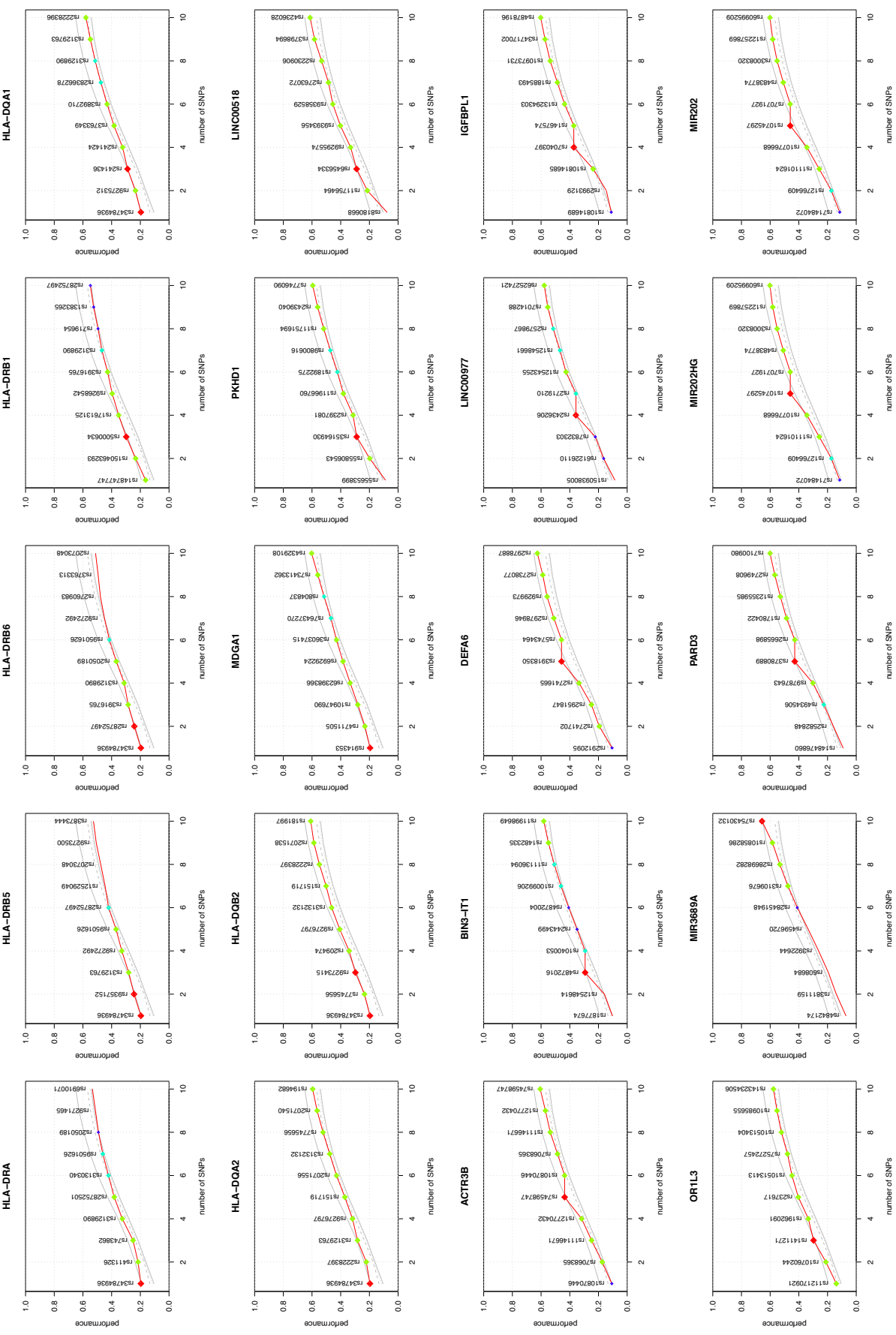

Figure A.1 b: Top genes - chromosome 6 to 10. Red denoted SNPs (◆) indicate a performance over 100 % of the extrapolated reference curve, SNPs marked green (◆) exceed 99 %, SNPs marked light blue (◆) 95 %, and dark blue colored SNPs (◆) indicate a performance over 90 % of the permutation results. SNPs below 90 % are not marked.

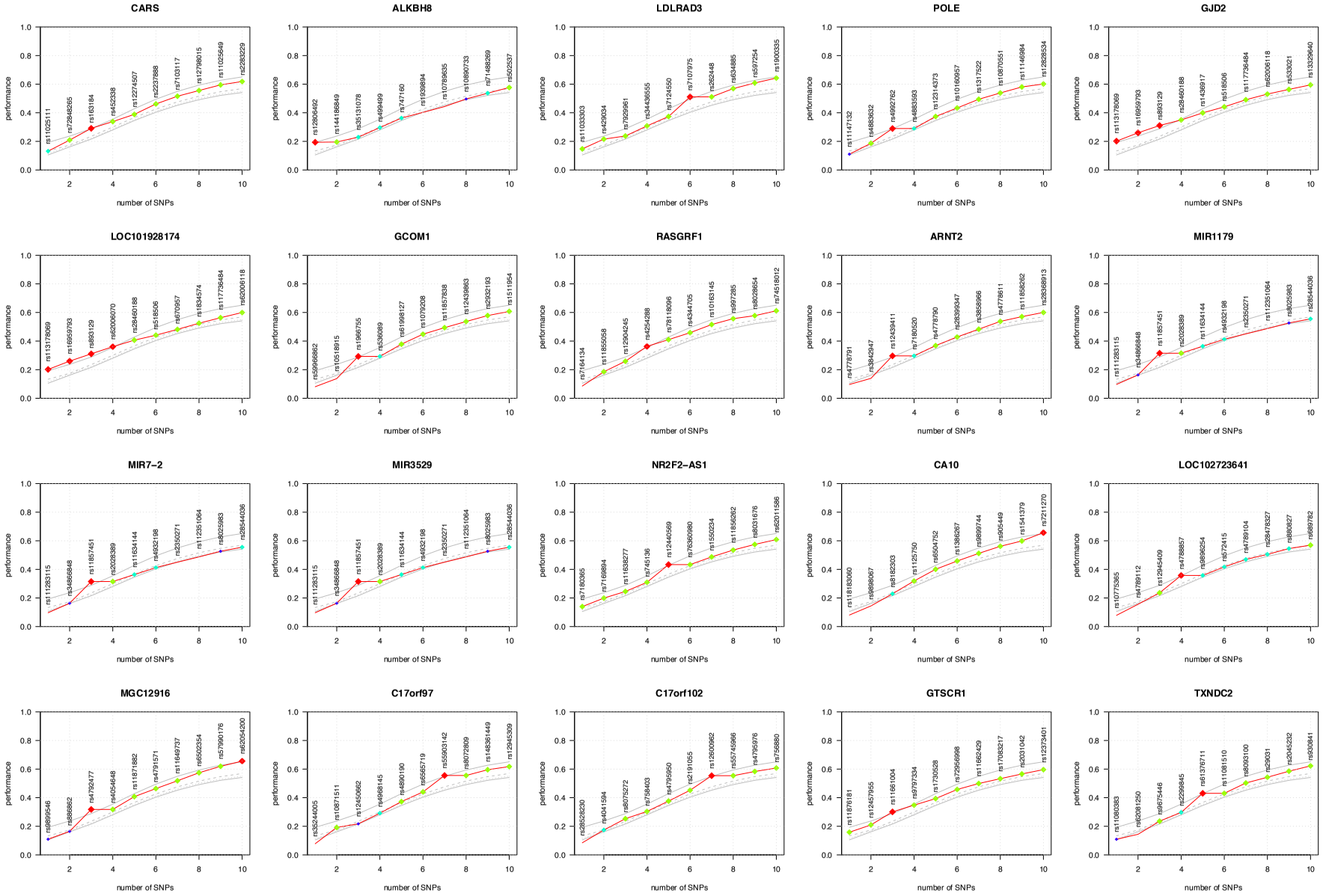

Figure A.1 c: Top genes - chromosome 11 to 18. Red denoted SNPs (◆) indicate a performance over 100 % of the extrapolated reference curve, SNPs marked green (◆) exceed 99 %, SNPs marked light blue (◆) 95 %, and dark blue colored SNPs (◆) indicate a performance over 90 % of the permutation results. SNPs below 90 % are not marked.

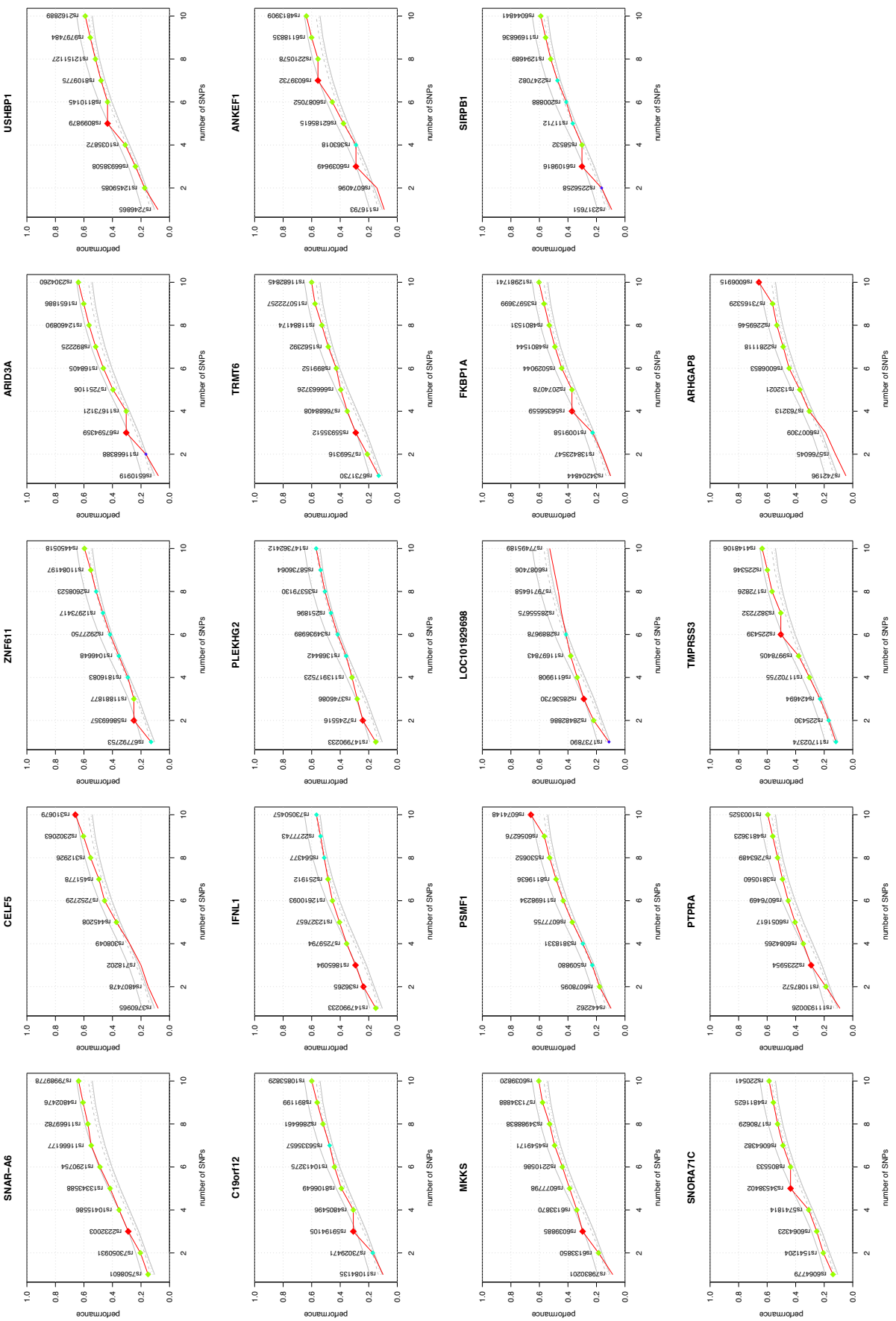

Figure A.1 d: Top genes - chromosome 19 to 22. Red denoted SNPs (◆) indicate a performance over 100 % of the extrapolated reference curve, SNPs marked green (◆) exceed 99 %, SNPs marked light blue (◆) 95 %, and dark blue colored SNPs (◆) indicate a performance over 90 % of the permutation results. SNPs below 90 % are not marked.

B Working procedures

In this following section, the working procedures will be presented. For this project, we used the package **e1071** Meyer (2012); Meyer et al. (2015) within the programming language **R** to perform calculations to create prediction models and detect the most relevant SNPs associated with antibody production against interferon-β.

For each genewise dataset, we performed the following steps:

The syntax of the `svm` call is as follows:

```
pf.data <- preFilter(data, ...)
```

The function is available in the package **caret** Kuhn et al. (2016). In the `findCorrelation()` function, the correlation threshold can be chosen. In this project we filtered all SNPs having a correlation >0.9.

```
prune <- prune.svm(formula, data, cost, gamma, ...)
```

The function `prune.svm()` performs the pruning calculations. For each step, new SVM models are created and evaluated. The parameters `cost` and `gamma` can be set within this command.

To review the significance of the results, we created a reference curve by permutating the phenotype. The rearranged data of randomly assigned phenotypes to different individuals represents the null-hypothesis. For each chromosome, I selected 5 random SNPs to be included for calculations creating the reference curve.

```
reference.curve <- perm.and.perf(data, m, ...)
```

In each iterations of the function `perm.and.perf()`, the phenotype was permutated, 10 of 110 included SNPs were selected, and the performance from 1 to 10 SNPs included in the pruning calculation was recorded. The iteration was repeated 1 million times, as indicated by the parameter `m`.

To identify which genes and relevant SNPs reach a higher performance compared to the referance curve, we wrote the function `sig.prune()`. The input genotype data and reference data needs to be provided. From this process, we received a list of the significant gene names.

```
significant.SNPs <- sig.prune(data, ref.curve, ...)
```

To obtain a final model which can be applied for multiple sclerosis patients to predict antibody prodution, we performed one more calculation. For this, all SNPs which exceeded the reference curve were summarized. This dataset, including 315 significant SNPs localized in 78 genomewide genes, was used to perform a `prune.svm()` calculation.

C Acknowledgements

Primarily and most of all I would like to express my sincere gratitude to my doctorate supervisor Prof. Dr. Bertram Müller-Myhsok. He kindly welcomed me to be part of his research group in Statistical Genetics at the Max-Planck Institute of Psychiatry in Munich and thoughtfully guided me through this project on genotype-based machine learning.

Secondly, I would like to thank Dr. Benno Pütz, who taught me the fundamental knowledge of programming in computer language and introduced me to machine learning approaches. He continuously assisted me in informatics and computational questions throughout this project.

I was very fortunate to be part of this fantastic research group and would like to hereby thank the entire team for all their support, assistance and input.

My sincere thanks also goes to the entire Max-Planck Institute of Psychiatry in Munich. I feel honored to have had this phenomenal opportunity to complete my doctoral thesis at this outstanding institution and contribute to exceptional research.

Furthermore, I would also like to gratefully thank two former group members Dr. rer. nat. Christiane Wolf and Dr. rer. nat. Darina Czamara, who assisted me kindly in initial working steps and data preparation.

I like to pronounce my sincere gratitude to the Department of Neurology at the Rechts der Isar Hospital. This shall apply particularly to Prof. Dr. med. Bernhard Hemmer and PD Dr. Dorothea Buck, who kindly provided us with the data on multiple sclerosis patients utilized in this thesis. I am very grateful for their supportive cooperation at all times. My gratitude is also due to all the patients, who generously consented to contribute to research.

My sincere thanks goes to Dr. Philipp Sämann and the department of magnetic resonance imaging at the Max-Planck Institute of Psychiatry in Munich, Germany for the courtesy of providing sample images, which display typical multiple sclerosis lesions.

I would like to express my special appreciation to all my friends, fellow students and professors at the Technical University of Munich, who accompanied and supported me during the course of completing this thesis, my studies at medical school and my years in Munich.

Finally but most lovingly, I want to thank my parents Dr. med. Sieglinde & Johannes Schmiedlechner and my siblings Patricia & Alexander for their infinite support, relentless motivation and loving care.

Bibliography

Association of British Pharmaceutical Industry, ABPI (n.y.). [The main forms of multiple sclerosis (MS)] Fingolimod - Novel Therapy for Multiple Sclerosis. *drugdevelopment-technology.com*. Image retrieved 28.04.2014.

Baranzini, S. (2011). Revealing the genetic basis of multiple sclerosis: Are we there yet? *Current Opinion in Genetics and Development*, 21:317–324.

Barbosa, M., Vielmetter, J., Chu, S., Smith, D., and Jacinto, J. (2006). Clinical link between MHC class II haplotype and interferon-beta (IFN-beta) immunogenicity. *Clinical Immunology*, 118:42–50.

Barsh, G., Copenhaver, G., Gibson, G., and Williams, S. (2012). Guidelines for Genome-Wide Association Studies. *PLOS Genetics*, 8.

Becker, N., Werft, W., and Benner, A. (2012). *penalizedSVM: Feature Selection SVM using penalty functions*. R package version 1.1. *https://CRAN.R-project.org/package=penalizedSVM*.

Becker, N., Werft, W., Toedt, G., Lichter, P., and Benner, A. (2009). Penalized SVM — an R package for feature selection SVM classification. *Bioinformatics*, 25:1711–1712.

Bennett, K. and Campbell, C. (2000). Support Vector Machines: Hype or Hallelujah? *SIGKDD Explorations*, 2:1–13.

Bethesda (2005). Database of Single Nucleotide Polymorphisms (dbSNP). National Center for Biotechnology Information, National Library of Medicine. *www.ncbi.nlm.nih.gov*.

Buck, D., Cepok, S., Hoffmann, S., Grummel, V., Jochim, A., Berthele, A., Hartung, H., Wassmuth, R., and Hemmer, B. (2011). Influence of the HLA-DRB1 Genotype on Antibody Development to Interferon Beta in Multiple Sclerosis. *Arch Neurol.*, 68(4):480–487.

Buck, D. and Hemmer, B. (2014). Biomarkers of treatment response in multiple sclerosis. *Expert Reviews Neurotherapy*, 14(2):165–172.

Cantor-Rivera, D., Khan, A., Goubran, M., Mirsattari, S., and Peters, T. (2014). Detection of temporal lobe epilepsy using support vector machines in multi-parametric quantitative MR imaging. *Computerized Medical Imaging and Graphics*, 41:14–28.

Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at *www.csie.ntu.edu.tw/~cjlin/libsvm*.

Comabella, M., Fernández-Arquero, M., Río, J., Guinea, A., Fernández, M., Cenit, M., de la Concha, E., and Montalban, X. (2009). HLA class I and II alleles and response to treatment with interferon-beta in relapsing-remitting multiple sclerosis. *Journal of Neuroimmunology*, 210:116–119.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S., and Tiffin, N. (2016). *RPostgreSQL: R interface to the PostgreSQL database system*. R package version 0.4-1. *https://CRAN.R-project.org/package=RPostgreSQL*.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Creeke, P. and Farrell, R. (2013). Clinical testing for neutralizing antibodies to interferon-β in multiple sclerosis. *Therapeutic Advances in Neurological Disorders*, 6:3–17.

Delaneau, O., Marchini, J., and Zagury, J. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9:179–181.

Delaneau, O., Zagury, J., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10:5–6.

DTREG - predictive modeling software (2014). [The Kernel trick] Introduction to Support Vector Machine (SVM) Models. *www.dtreg.com*. Image retrieved 14.05.2014.

Enevold, C., Oturai, A., Soelberg Sorensen, P., Ryder, L., Koch-Henriksen, N., and Bendtzen, K. (2010). Polymorphisms of innate pattern recognition receptors, response to interferon-beta and development of neutralizing antibodies in multiple sclerosis patients. *Multiple Sclerosis Journal*, 16(8):942–949.

Freeman, C. and Marchini, J. (2012). GTOOL. Version 0.7.5. Software available at *www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html*.

Glessner, J., Wang, K., Sleiman, P., Zhang, H., Kim, C., Flory, J., Bradfield, J., Imielinski, M., Frackelton, E., Qiu, H., Mentch, F., Grant, S., and Hakonarson, H. (2010). Duplication of the SLIT3 Locus on 5q35.1 Predisposes to Major Depressive Disorder. *PLoS One Public Library of Science*, 5.

Grehl, H. and Reinhardt, F. (2013). *Checkliste Neurologie*, volume 5. Georg Thieme Verlag.

Guipponi, M., Vuagniaux, G., Wattenhofer, M., Shibuya, K., Vazquez, M., Dougherty, L., Scamuffa, N., Guida, E., Okui, M., Rossier, C., Hancock, M., Buchet, K., Reymond, A., Hummler, E., Marzella, P., Kudoh, J., Shimizu, N., Scott, H., Antonarakis, S., and Rossier, B. (2002). The transmembrane serine protease (TMPRSS3) mutated in deafness DFNB8/10 activates the epithelial sodium channel (ENaC) in vitro. *Human Molecular Genetics*, 11(23):2829–2836.

Hafler, D. (2004). Multiple Sclerosis. *Journal of Clinical Investigation*, 113(6):788–794.

Hartung, H., Haas, J., Meergans, M., Tracik, F., and Ortler, S. (2013). Interferon-β_{1b} in der Multiple Sklerose Therapie, mehr als 20 Jahre klinische Erfahrung. *Nervenarzt*, 84(6):679–704.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, New York, 2nd edition.

Hoffmann, S., Cepok, S., Grummel, V., Lehmann-Horn, K., Hackermueller, J., Stadler, P., Hartung, H., Berthele, A., Deisenhammer, F., Wasmuth, R., and Hemmer, B. (2008). HLA-DRB1*0401 and HLA-DRB1*0408 Are Strongly Associated with the Development of Antibodies against Interferon-β Therapy in Multiple Sclerosis. *The American Journal of Human Genetics*, 83:219–227.

Howie, B., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6).

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3 Journal*, 1(6):457–470.

International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham, A., Patsopoulos, N., Xifara, D., Davis, M., Kemppinen, A., Cotsapas, C., Shahi, T., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., Fontaine, B., Hemmer, B., Martin, C., Zipp, F., D'alfonso, S., Martinelli-Boneschi, F., Taylor, B., Harbo, H., Kockum, I., Hillert, J., Olsson, T., Ban, M., Oksenberg, J., Hintzen, R., Barcellos, L., Wellcome Trust Case Control Consortium 2 (WTCCC2), International IBD Genetics Consortium (IIBDGC), Agliardi, C., Alfredsson, L., Alizadeh, M., Anderson, C., Andrews, R., Sondergaard, H., Baker, A., Band, G., Baranzini, S., Barizzone, N., Barrett, J., Bellenguez, C., Bergamaschi, L., Bernardinelli, L., Berthele, A., Biberacher, V., Binder, T., Blackburn, H., Bomfim, I., Brambilla, P., Broadley, S., Brochet, B., Brundin, L., Buck, D., Butzkueven, H., Caillier, S., Camu, W., Carpentier, W., Cavalla, P., Celius, E., Coman, I., Comi, G., Corrado, L., Cosemans, L., Cournu-Rebeix, I.and Cree, B., Cusi, D., Damotte, V., Defer, G., Delgado, S., Deloukas, P., di Sapio, A., Dilthey, A., Donnelly, P., Dubois, B., Duddy, M., Edkins, S., Elovaara, I., Esposito, F., Evangelou, N., Fiddes, B., Field, J., Franke, A., Freeman, C., Frohlich, I., Galimberti, D., Gieger, C., Gourraud, P., Graetz, C., Graham, A., Grummel, V., Guaschino, C., Hadjixenofontos, A., Hakonarson, H., Halfpenny, C., Hall, G., Hall, P., Hamsten, A., Harley, J., Harrower, T., Hawkins, C., Hellenthal, G., Hillier, C., Hobart, J., Hoshi, M., Hunt, S., Jagodic, M., Jelcic, I., Jochim, A., Kendall, B., Kermode, A., Kilpatrick, T., Koivisto, K., Konidari, I., Korn, T., Kronsbein, H., Langford, C., Larsson, M., Lathrop, M., Lebrun-Frenay, C., Lechner-Scott, J., Lee, M., Leone, M., Leppä, V., Liberatore, G., Lie, B., Lill, C., Lindén, M., Link, J., Luessi, F., Lycke, J., Macciardi, F., Männistö, S.and Manrique, C., Martin, R., Martinelli, V., Mason, D., Mazibrada, G., McCabe, C., Mero, I., Mescheriakova, J., Moutsianas, L., Myhr, K., Nagels, G., Nicholas, R., Nilsson, P., Piehl, F., Pirinen, M., Price, S., Quach, H., Reunanen, M., Robberecht, W., Robertson, N., Rodegher, M., Rog, D., Salvetti, M., Schnetz-Boutaud, N., Sellebjerg, F., Selter, R., Schaefer, C., Shaunak, S., Shen, L., Shields, S.,

Siffrin, V., Slee, M., Soelberg Sorensen, P., Sorosina, M., Sospedra, M., Spurkland, A., Strange, A., Sundqvist, E., Thijs, V., Thorpe, J., Ticca, A., Tienari, P., van Duijn, C., Visser, E., Vucic, S., Westerlind, H., Wiley, J., Wilkins, A., Wilson, J., Winkelmann, J., Zajicek, J., Zindler, E., Haines, J., Pericak-Vance, M., Ivinson, A., Stewart, G., Hafler, D., Hauser, S., Compston, A., McVean, G., De Jager, P., Sawcer, S., and McCauley, J. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11):1353–1362. Image retrieved 19.08.2014.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Karatzoglou, A., Smola, A., and Hornik, K. (2016). *kernlab: Kernel-Based Machine Learning Lab*. R package version 0.9-25. *https://CRAN.R-project.org/package=kernlab*.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., and Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006.

Kivisäkk, P., Alm, G., Tian, W., Matusevicius, D., Fredrikson, S., and Link, H. (1997). Neutralising and binding anti-interferon-β–Ib (IFN-β–Ib) antibodies during IFN-β–Ib treatment of multiple sclerosis. *Multiple Sclerosis Journal*, 3:184–190.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt, T. (2016). *caret: Classification and Regression Training*. R package version 6.0-73. *https://CRAN.R-project.org/package=caret*.

Li, Y., Pabst, S., Lokhande, S., Grohe, C., and Wollnik, B. (2009). Extended genetic analysis of BTNL2 in sarcoidosis. *Tissue Antigens*, 73:59–61.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Liaw, A. and Wiener, M. (2015). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-12. *https://CRAN.R-project.org/package=randomForest*.

Link, J., Lundkvist Ryner, M., Fink, K., Hermanrud, C., Lima, I., Brynedal, B., Kockum, I., Hillert, J., and Fogdell-Hahn, A. (2014). Human Leukocyte Antigen Genes and Interferon Beta Preparations Influence Risk of Developing Neutralizing Anti-Drug Antibodies in Multiple Sclerosis. *PLoS One Public Library of Science*, 9:e90479.

Meyer, D. (2012). *Support Vector Machines* The Interface to `libsvm` in package e1071*. Technische Universität Wien, Austria, David.Meyer@ci.tuwien.ac.at. *http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf*.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., and Lin, C. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. *https://CRAN.R-project.org/package=e1071*.

Mi, H., Muruganujan, A., Casagrande, J., and Thomas, P. (2013a). PANTHER Classification System - Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, pages 1551 – 1566.

Mi, H., Muruganujan, A., and Thomas, P. (2013b). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, pages 377–386.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J., and Thomas, P. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, 44:336–342.

Mi, H. and Thomas, P. (2009). PANTHER pathway - Protein Networks and Pathway Analysis. *Methods in Molecular Biology*, 563:123–140.

Morais, A., Lima, B., Peixoto, M., Alves, H., Marques, A., and Delgado, L. (2012). BTNL2 gene polymorphism associations with susceptibility and phenotype expression in sarcoidosis. *Respiratory Medicine*, 106:1771–1777.

Müller, K., Wickham, H., James, D., Falcon, S., SQLite Authors, Healy, L., R Consortium, and RStudio (2017). *RSQLite: 'SQLite' Interface for R*. R package version 1.1-2. *https://CRAN.R-project.org/package=RSQLite*.

Multiple Sclerosis International Federation (2013). [Prevalence of multiple sclerosis] ATLAS OF MS 2013: MAPPING MULTIPLE SCLEROSIS AROUND THE WORLD. *www.msif.org/wp-content/uploads/2014/09/Atlas-of-MS.pdf*. Image retrieved 30.04.2014.

Purcell, S. (2006). PLINK version 1.07. *http://pngu.mgh.harvard.edu/purcell/plink/*.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., and Sham, P. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81(3):559–575.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. R Foundation for Statistical Computing *http://www.R-project.org*.

Runne, C. and Chen, S. (2013). PLEKHG2 Promotes Heterotrimeric G Protein βγ—Stimulated Lymphocyte Migration via Rac and Cdc42 Activation and Actin Polymerization. *Molecular and Cellular Biology*, 33:4294–4307.

Sato, K., Sugiyama, T., Nagase, T., Kitade, Y., and Ueda, H. (2014). Threonine 680 phosphorylation of FLJ00018/PLEKHG2, a Rho family-specific guanine nucleotide exchange factor, by epidermal growth factor receptor signaling regulates cell morphology of Neuro-2a cells. *Journal of Biological Chemistry*, 289:10045–56.

Schölkopf, B., Smola, A., and Müller, K. (1997). Kernel principal component analysis. In *Artificial Neural Networks — ICANN '97*, volume 1327 of *Lecture Notes in Computer Science*, pages 583–588.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Sitzer, M. and Steinmetz, H. (2011). *Lehrbuch Neurologie*. Elsevier, Urban und Fischer, 1st edition.

Smola, A., Schölkopf, B., and Müller, K. (1998). The connection between regularization operators and support vector kernels. *Neural Networks*, 11(4):637–649.

Soelberg Sorensen, P. (2008). Neutralizing Antibodies Against Interferon-Beta. *Therapeutic Advances in Neurological Disorders*, 1:62–78.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526:68/74.

The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426:789–796.

The International Multiple Sclerosis Genetics Consortium and the Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476:214–219.

Toshimoto, K., Wakayama, N., Kusama, M., Maeda, K., Sugiyama, Y., and Akiyama, Y. (2014). In silico Prediction of Major Drug Clearance Pathways by Support Vector Machines with Feature-selected Descriptors. *Drug Metabolism and Disposition*, 42:1811–1819.

Wallrapp, C., Hähnel, S., Müller-Pillasch, F., Burghardt, B., Iwamura, T., Ruthenbürger, M., Lerch, M., Adler, G., and Gress, T. (2000). A Novel Transmembrane Serine Protease (TMPRSS3) Overexpressed in Pancreatic Cancer. *Cancer Research*, 60(10):2602–2606.

Weber, F., Cepok, S., Wolf, C., Berthele, A., Uhr, M., Bettecken, T., Buck, D., Hartung, H., Holsboer, F., Müller-Myhsok, B., and Hemmer, B. (2012). Single-nucleotide polymorphism in HLA- and non-HLA genes associated with the development of antibodies to interferon-β therapy in multiple sclerosis patients. *The Pharmacogenomics Journal*, 12:238–245.

Wennerstroem, A., Pietinalho, A., Lasota, J., Salli, K., Surakka, I., Seppaenen, M., Selroos, O., and Lokki, M. (2013). Major histocompatibility complex class II and BTNL2 associations in sarcoidosis. *European Respiratory Journal*, 42:550–553.

Index

Index of genes

Index of SNPs